MASTER OF SCIENCE BY RESEARCH

Quantitative analysis of evolution and role behaviour in an online social network

Hitchinson, Eleni

Award date:
2013

Awarding institution:
Coventry University

Coventry University

Faculty of Engineering and Computing

QUANTITATIVE ANALYSIS OF EVOLUTION AND ROLE BEHAVIOUR IN AN ONLINE SOCIAL NETWORK

Eleni Hitchinson

Submitted for Master by Research (MScR), April 2013

Contents Page

Figures

Tables

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank all those who have helped me over the past two years to produce this report. This report was only possible due to the common interests of many people. I would like to especially thank my supervisor Dr Christian von Ferber, who has given me the opportunity, advice and support throughout.

I would also like to thank my family and friends who have supported and motivated me with my project.

Thank you

ABSTRACT

A number of online discussion groups have a long history where individual users are found to participate over long time ranges. These groups therefore offer the possibility to test hypothesis such as preferential attachment on such time scales. The focus of this thesis is in particular to develop quantitative indicators for the type of discussion (e.g. philosophical or technical) and the self-defined roles of the participants, [Chang2002].

Investigations into these two groups confirm similarities and differences in statistical properties of the networks. The degree distribution, network size, clustering and betweeness are all examined. New measures introduced, include the reply count and positions of the posts and globally each group is compared to each other.

Top actors of both groups are selected exploring their individual networks, through the use of Gephi an open source graphical manipulation software, [Bastian2009].

Through analysing the discussions three roles are observed, the answer role, question role and discussion role. Developing indicators for these roles observe quantitatively how these roles are classified.

INTRODUCTION

Networks are the mathematical make up of the framework by which we live. They can be as complex as the transfer of data from computer to computer, the interactions between proteins or as simple as catching a bus to school. Online social networks are the most popular sites on the World Wide Web, many of which are highly marketable, thus investigation into posting frequency and particular behaviour of 0the high frequency posters within these sites is may be expected to reveal interesting insight.

Networks are displayed using the basis of graph theory, which has been widely researched since the 17th century when the Königsberg bridge problem was solved by Euler and is now accepted as the first proof of network theory. However it was not until early in the 21st century that the popularity of network research boomed, one of the earlier studies was that of school children (Wellman, 1926, as cited by Boccaletti2006). Not only were the scientist's and mathematicians interested but also sociologists, as networks can help explain the friendship network behaviour of relationships between people. Originally only small networks were explored but now the statistical behaviour of large-scale real world networks is commonly researched.

The Cambridge Dictionary states a network *'is a large system consisting of many similar parts that are connected together to allow movement or communication between or along the parts or between the parts and a control centre'.* This general

description can be used to model any real complex network, [Albert2002] [Boccaletti2006] [Newman2003]

Common properties of online social networks are provided along with the common roles found within these networks. The project investigates two very different discussion groups to find quantitative indicators for these social roles. Methods are introduced and developed to identify these roles, [Hannemann2005].

1.0 Properties of social networks

To fully understand the behaviour of individuals in groups and of groups as a whole, insight into the mathematical properties of the network is required. These are general network properties, which have been widely researched over many decades. In the following, we develop these concepts for the special case of online discussion group networks, [Hannemann2005][Newman2002].

Here, actor is the term used to describe the individual user within the network, be it that he is posting or replying to a discussion. Vertices and edges are displayed in graphs to show the structure of a discussion, with vertices, N, representing, the actor or the post and edges, E, showing the relationship between two actors or posts (figure 1). Please note that a vertex may equivalently also be called a 'node'.

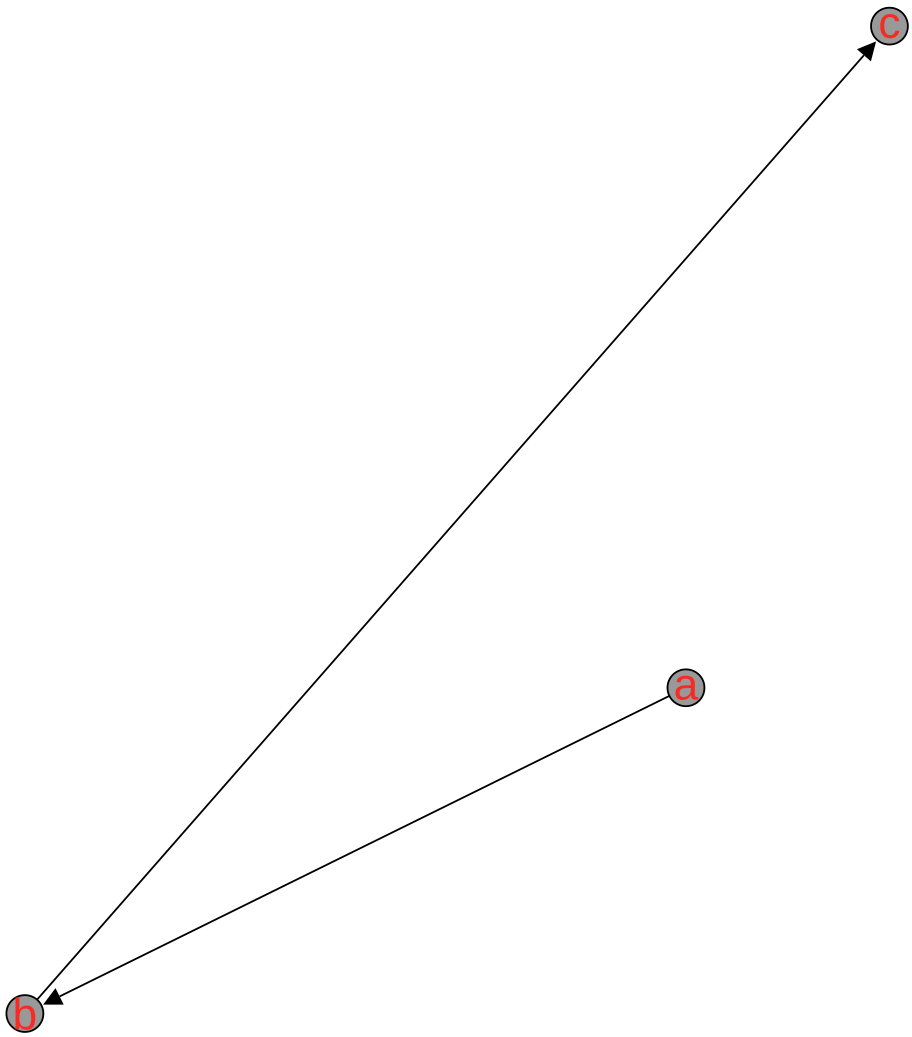


Figure 1: Network between three actors

As the data within this paper will be formed using both directed and undirected data, properties of both types will be explained.

Two vertices are joined together by an edge to show that there is a tie between them. Within a directed network this edge will be directed, i.e. the line will have an arrowhead pointing to one or both vertices. An example of this would be an email between three acquaintances, Mark, Claire and John. If Claire sends John an email

then there would be an arrow from Claire to John directed towards John. If John sends Mark an email there will be a directed edge from John to Mark. In an undirected network there would be a straight line with no arrowheads. These vertices and edges combine to form a network or social graph.

It is also possible for the directed tie to have two arrow heads, one in each direction, in the example if John replies to Claire's email then the edge between will consist of an arrow head at both ends (figure 2) this would show a reciprocated, mutual and co-occurring relationship between them both.

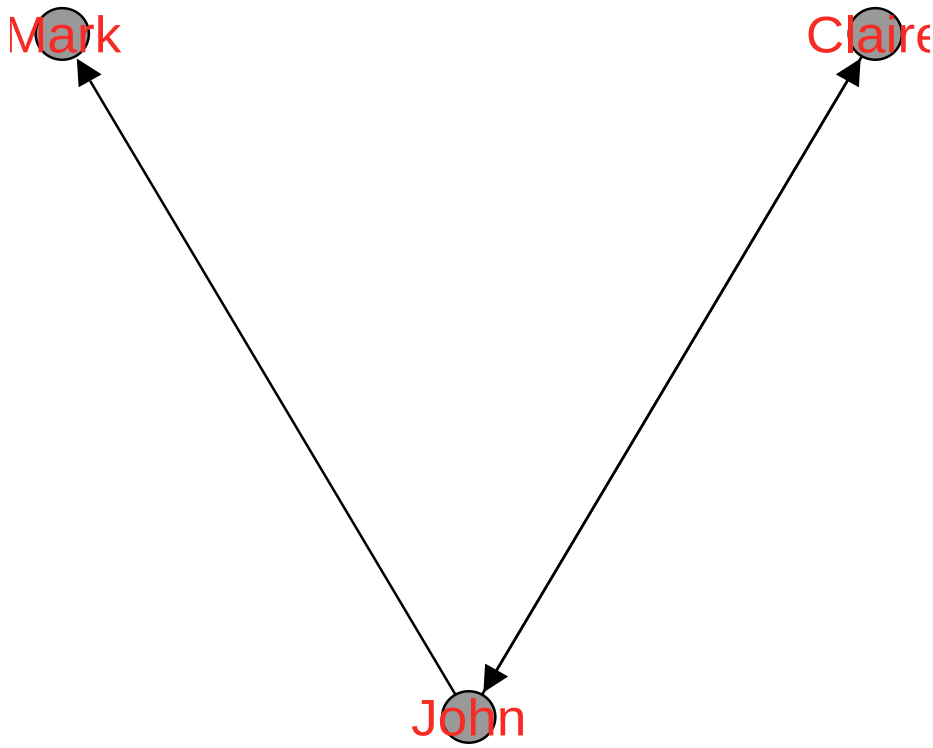


Figure 2 : Directed Edges

An edge can also be weighted, using the example between two acquaintances if Claire emails John three times, there will be one connection but it will carry a

weight of three. If John only emails Claire once there will be a connection carrying a weight of one.

1.1 The Degree

One of the most important properties of the vertices within a network is the degree, denoted as k, the number of connections (edges) a vertex has with other vertices [Dorogovtsev2002]

Within a directed network, the degree is split into two, an in-degree and an out-degree. The in-degree is the number of in-coming edges that a vertex has directed towards it and the out-degree is the number of out-going edges directed away from it. It is useful to treat these as different relationships formed by the in and out-degree. Thus the total degree of a single directed vertex will be the sum of the in-degree and out-degree at this vertex (equation 1.1.1, figure 3).

The total degree of the whole network of a directed graph is the number of connections formed between vertices. The summation of the in-degree and out-degree results in the total degree.

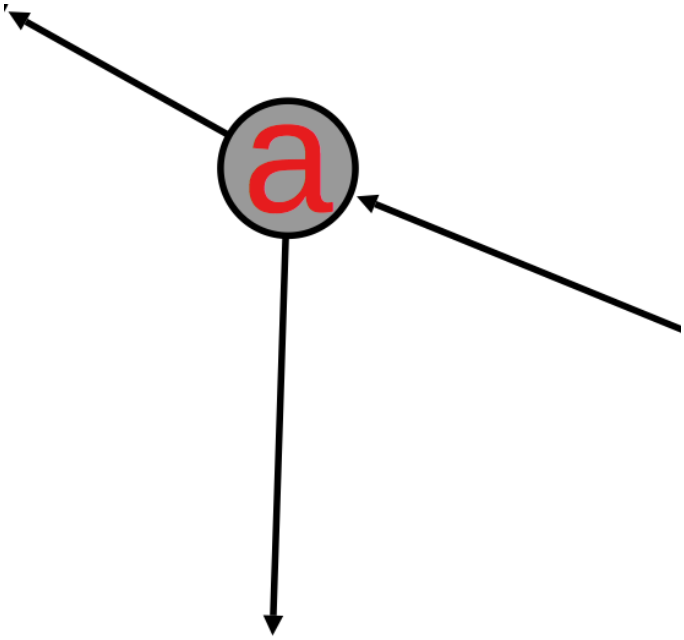


Figure 3 : The in-degree and out-degree of a vertex

$$k_i = 1$$

$$k_o = 2$$

$$k = k_i + k_o \qquad (1.1.1)$$

$$k = 3$$

The average degree, $\bar{k}$, of the network is the sum of all degrees divided by the total number $N$of vertices within the network.

$$\bar{k} = \frac{\sum_i k_i}{N} \qquad (1.1.2)$$

Vertices with high total degree, high in-degree and high out-degree can be easily identified and provide great detail about the individual actors and the group.

1.2 The Degree Distribution

The degree distribution counts the number of vertices that have a given degree. It is usually written as the normalised probability that a randomly chosen vertex, s, within a network of, N, vertices has exactly k amount of edges.

$$p(k, s, N) \quad k > 0 \qquad (1.2.1)$$

[Dorogovtsev2002]

For a directed network, the distributions of the in-degree and out-degree are respectively

$$p^i(k_i, s, N) \quad (1.2.2)$$

$$p^o(k_o, s, N) \; . \; (1.2.3)$$

By adding all the degree distributions of all the vertices within the network of size N gives the total degree distribution $P\,(k, N)$ of the whole network as shown in equation 1.2.4.

$$P(k, N) = \frac{1}{N} \sum_{s=1}^{N} p(k, s, N) \qquad (1.2.4)$$

There is a fundamental relation between the number of edges and the sum of all degrees: as far as each edge contributes to the degree of both vertices that it connects, the sum of all degrees k_i must be equal to twice the number of edges E, for a non-directed, non-weighted network.

$$2E = \sum_i k_i = N\bar{k} \qquad (1.2.4)$$

where N is the number of vertices and $\bar{k}$ is the mean degree.

The degree distribution alone can be very noisy and is often hard to follow on a histogram. By using the cumulative degree distribution clearer results may be obtained. The cumulative degree distribution, $P_{cum}(k)$, is defined as the probability of any vertex having a degree greater than or equal to k, and so $P_{cum}(k) = \sum_{k'=k}^{\infty} P(k')$. Due to the definition of the cumulative distribution the slope will always be decreasing. The character of the slope may be used to classify the distribution. In figure 4 the slope is linear in the double logarithmic plot. This indicates that the distribution decays according to a power law $P_{cum}(k) \sim k^{-\gamma}$.

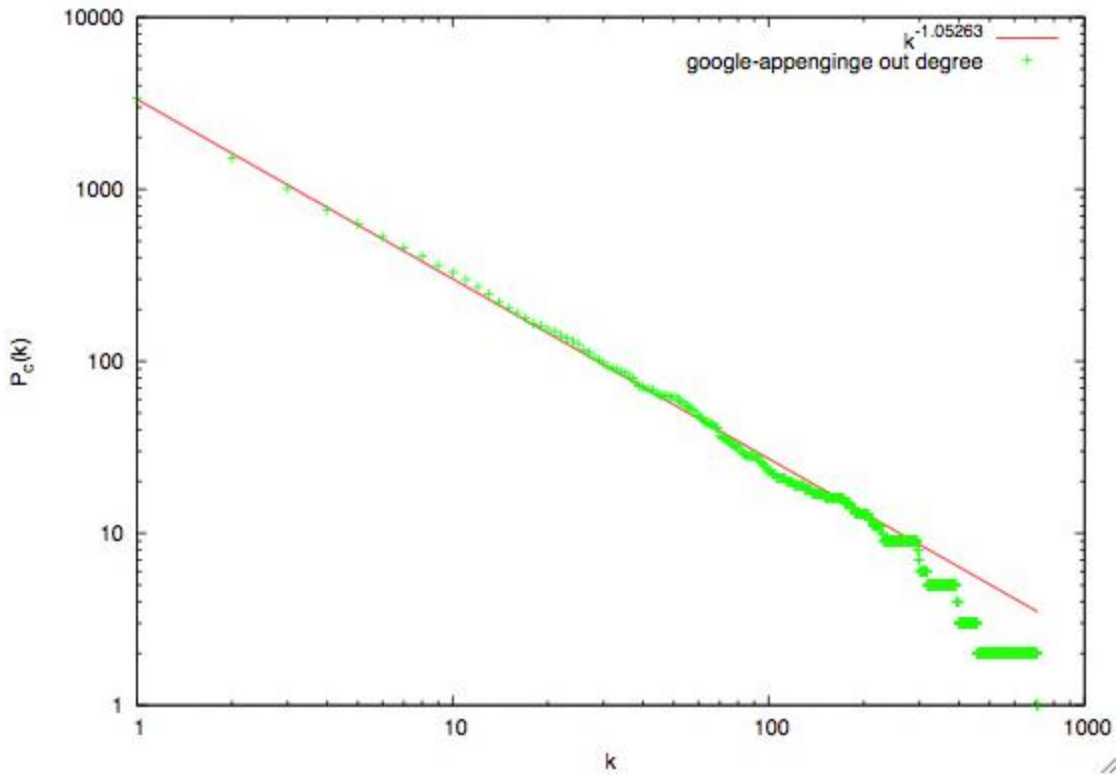


Figure 4 : Cumulative degree distribution P(k) of the discussion group

Other types of distribution include the poisson distribution, the exponential distribution, the power-law distribution and the multifractal distribution.

1.3 Network Size

Many properties, the number of actors, and the number of connections between actors, the average shortest path length, and the diameter may be used to define the size of the network. The size is critical for the development, maintenance and security of the network.

1.3.1 The Path Length

The shortest path length describes the distance from one vertex to another in terms of the minimal number of edges to be traversed between them. The average shortest path length is the average path length calculated from averaging the paths lengths between all pairs of connected vertices *a* and *b* of the graph.

In an undirected network this is a simple manoeuvre, as shown in figure 5. The shortest average path can be complex in a directed net, as shown in figure 6. A route from vertex *a* to vertex *b* may be completely different to that from vertex *b* to vertex *a* or even in some cases just because one may be able to reach *b* from *a* does not necessarily mean *a* can be reached from *b*.

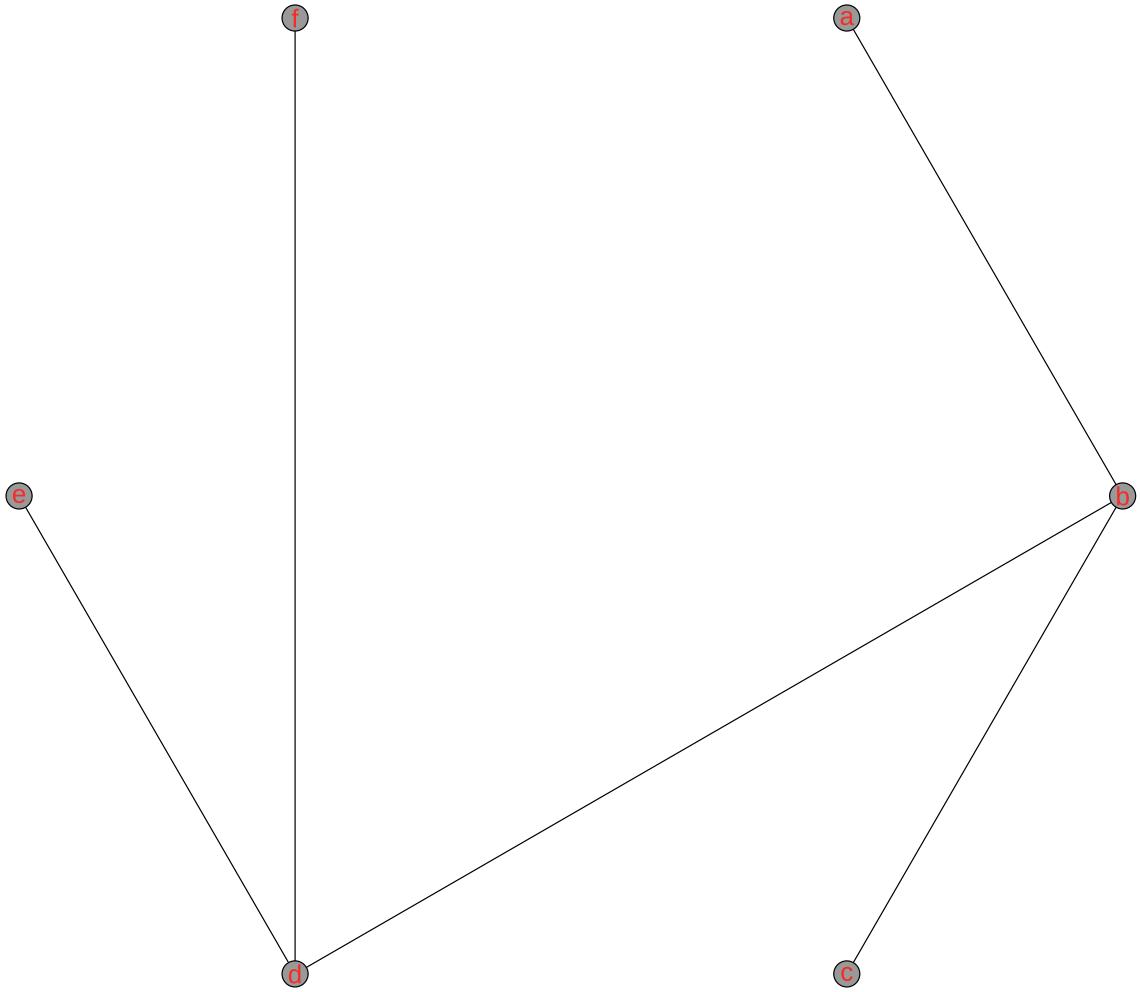


Figure 4 : The average shortest path length in an undirected network, from *a* to *c* is 2

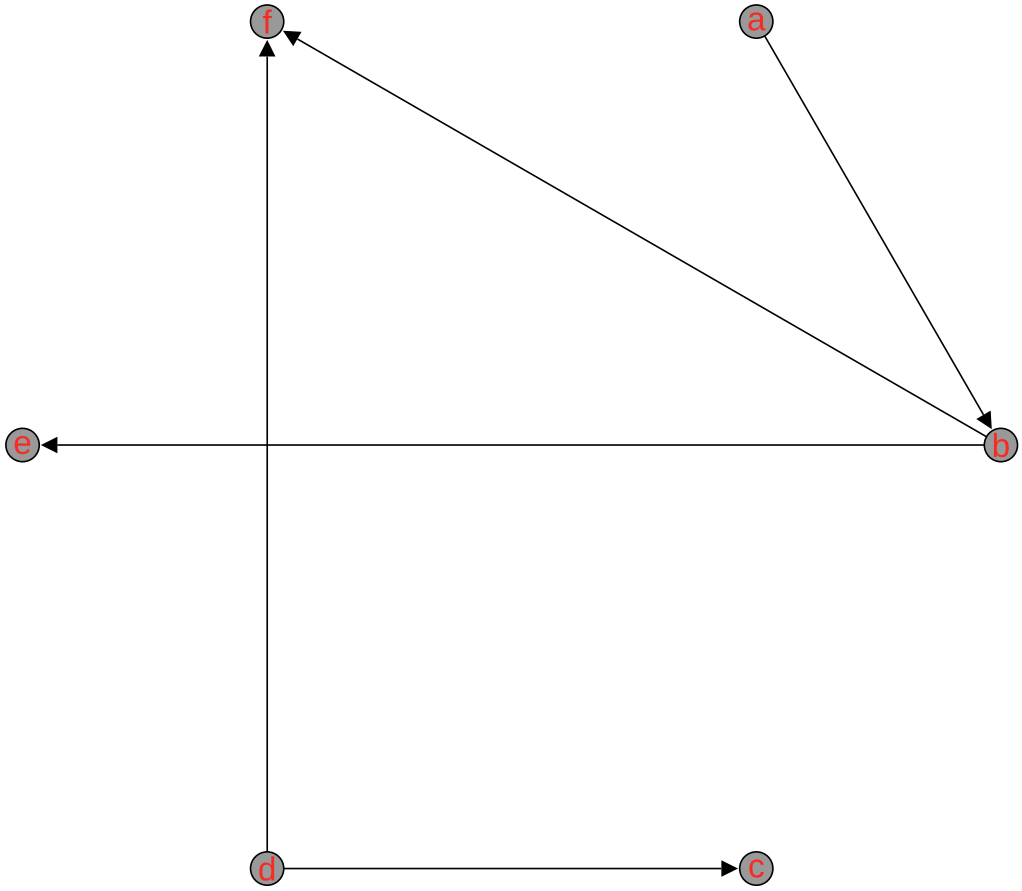


Figure 5 : The average shortest path length of a directed network.

A path length of one means only one edge links the two vertices and hence are nearest neighbours of one another.

The average shortest path length of a network describes how well connected the actors are within the network. Large dense networks are likely to display shorter average path lengths, as a high number of connections are present. A large sparse network is likely to display a higher average shortest path length, as there are fewer connections between any two vertices.

The average shortest path length of the network is calculated as the average of the shortest paths between all pairs of vertices in the network [Boccaletti2006] for a non-directed network this reads

$$\bar{l} = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{i-1} l_{min}\,(ij) \quad (1.3.1.1)$$

Actors who can reach and be reached by other actors at short path lengths are in a prominent position within the network. We will quantify this property further below in terms of 'Centrality'.

The diameter, D, of the total network is the measurement of the maximum shortest path length between any two vertices.

$$D = \max._{ij}\, l_{min}(ij) \qquad (1.3.1.2)$$

The graph centrality of a vertex *i* is the inverse maximum shortest path from this vertex *i* to all other vertices *j* within the graph

$$C_i = \frac{1}{\max_j l_{min}(ij)} \quad (1.3.1.3)$$

1.3.2 Clustering coefficient

Watts, Strogatz and Newman, (2002) define the clustering coefficient, c_i of a particular vertex *i* in terms of the connections between neighbours, the probability that two vertices m and j are connected to one another if they are both connected to vertex i.

For an undirected network the clustering coefficient of a particular vertex a, can be calculated as the number of edges $y_a$that connect the nearest $k_a$neighbours (e.g. b, c and d) of the vertex a, divided by the maximum possible number $k_a(k_a - 1)/2$ of edges between the k_a neighbours of vertex a. The clustering coefficient has a value in the range of 0 and 1. A value of 1 indicates that all of the closest neighbours of the vertex are connected to form a complete subgraph.

$$c_a = \frac{2y_a}{k_a(k_a-1)} \qquad (1.3.2.1)$$

A vertex that displays a high clustering coefficient means that a large percentage of its nearest neighbours are connected between each other.

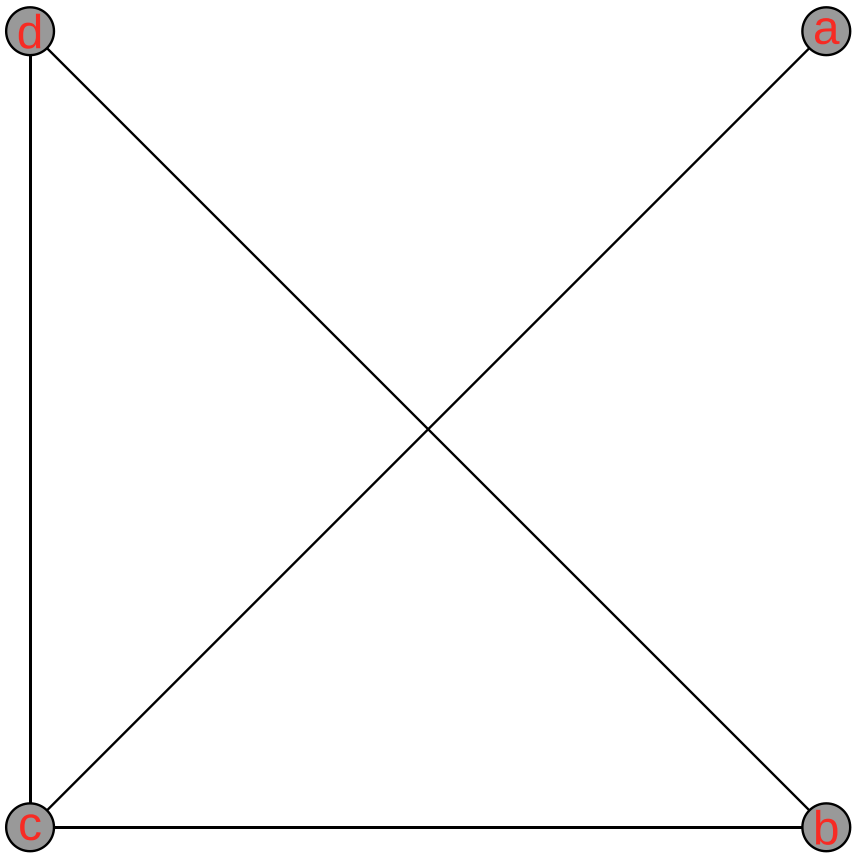


Figure 6: Undirected network

Vertex	y_a	k_a	c_a
A	0	1	0
B	1	2	1
C	1	3	0.333
D	1	2	1

Table 1 : Clustering coefficient is calculated for network in figure 7

For a directed network the clustering coefficient is produced on relationships between triads (3 actors), *a, b* and *c*. If there is a directed edge from *b* to *c*, will there be a directed edge from *a* to *c*?

$$C = \frac{no\ of\ transitive\ triads}{no.of\ potential\ transitive\ triads} \qquad (1.3.2.2)$$

The global average clustering coefficient of the entire network measures how well connected neighbours of any actor are. This is found by adding the clustering coefficients of all of the vertices then dividing by the total number of vertices within the whole network, equation no 1.3.2.3.

$$C = \frac{1}{N}\sum_i c_i \quad (1.3.2.3)$$

A network that displays a high global clustering coefficient is a highly connected network where the nearest neighbours of vertices have connections between each other.

1.3.3 Betweenness Centrality

Centrality measures describe the power particular individual actors have in their position within the network. They measure the importance of this individual in providing links and keeping the network connected, [Freeman1979].

Betweenness of a vertex measures the importance of a particular vertex to the connection between other vertices of the network. Figure 8 shows a path connecting vertices and to travel from *a* to *e* one would need to travel from *a* to *b*, *b*

to *c*, *c* to *d*, and *d* to *e*. This would be the shortest possible distance from *a* to *e*. There is no other possible route to get from *a* to *e*, thus *b, c,* and *d*, are all extremely important to the connectivity for this graph and hence if any of *b, c* and *d* were to be removed it would disconnect the graph and no path would exist from *a* to *e*. Betweenness is a measure of this importance, equation 1.3.3.1, [Boccaletti2006].

$$b_i = \sum_{j,k \epsilon N} \frac{n_{jk}(i)}{n_{jk}} \quad (1.3.3.1)$$

where n_{jk} is the number of shortest paths from *j* to *k* and $n_{jk}(i)$ is the number of shortest paths for travelling from *j* to *k* while passing through *i*

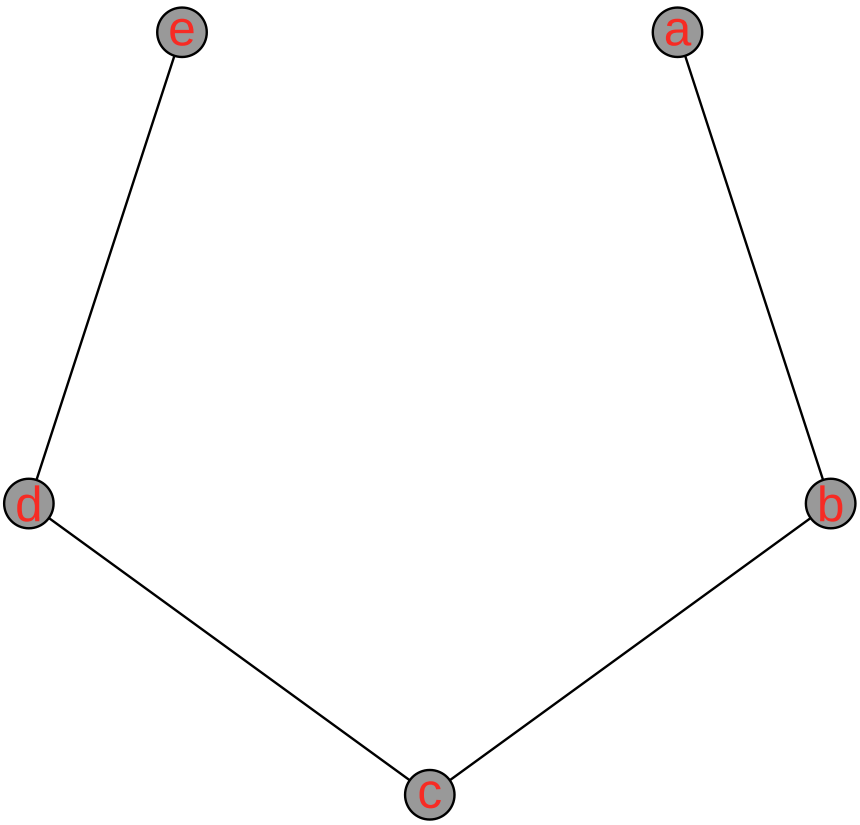


Figure 7 : Importance of vertices *b*, *c* and *d*

The betweenness of an edge *e* of a graph is similarly defined as the sum of shortest paths between all pairs of vertices *j* and *k* that pass through the edge *e* normalised by the overall number.

$$b_e = \sum_{j,k \epsilon N} \frac{n_{jk}(e)}{n_{jk}} \qquad (1.3.3.1)$$

Betweenness can be calculated on both directed and undirected graphs.

1.3.4 Degree Centrality

The degree centrality describes how an actor can be in a more prominent position if they have a higher total degree over an actor with a less degree. It is more advantageous for the actor with high degree as it provides greater opportunities and more choices. More connections to other actor's provide more choice of where to get information from. For example an actor who is connected to three other actors (in an undirected network) can choose to collect information from maybe one, two or even all three, but an actor with only one connection can only get information from this one connection.

A normalised degree centrality is calculated by

$$C_d = \frac{k}{N-1} \qquad (1.3.4.1)$$

where *N* is the number of vertices.

There are limitations to the degree centrality as it only focuses on the closest neighbour to the said actor and not the *m-th* nearest neighbour. This said an actor may have many connections and the actors they are connected to may have few connections themselves, where as an actor who has few connection to other actors may be in a more favourable position if the connected actors themselves have a large degree.

2.0 SCALE FREE DISTRIBUTIONS

Many real world networks are classified as scale-free. This section describes how such graphs are formed and evolve. The subject of preferential attachment is introduced in the following sub section.

Many real-world networks follow a power-law degree distribution, which differs from the poisson distribution in that the probability for a vertex to have a high degree k is for large k much higher than would result from the poisson distribution found for random networks. In these real world networks the degree distribution decays for large k following a power law, [Molloy1995].

$$P(k) \propto k^{-\gamma}, k \neq 0 \quad (2.0.1)$$

The exponent γ is often found to be between 2 and 3 but may also attain lower and higher values.

Networks with power law degree distributions are often called scale free due to the power-law having the property of having the same functional form at all scales and there is no scale present on which the distribution decays as for exponential or poisson distributions.

A network can be constructed through two different ways:

- Random Growth
- Growth with preferential attachment

2.1 Random Growth

If a new vertex s is introduced to a network, a by randomly attaching this vertex to an already existing vertex a then this increases the degree of the new vertex s and the existing vertex a, figure 9, [Dorogovtsev2002]

$$k_a \to k_a + 1$$

$$k_s \to k_s + 1$$

For a randomly growing network we assume that we start with a single vertex and at each unit time step, t, a new edge is added connecting a vertex s to a randomly chosen vertex. Thus the total degree becomes

$$\sum k = 2t$$

As the network grows the degree k of vertex s at time t, $k(s,t)$ evolves as follows. We assume t to be large and approximate the discrete growth by a differential equation:

$$\frac{\delta k(s,t)}{\delta t} = \frac{2}{t} \qquad (2.1.1)$$

$$k(s,t) = \int \delta k = 2\int \frac{\delta t}{t} + C(s)$$

$$k(s,t) = 2lnt + C(s),$$

where $C(s)$is the integration constant.

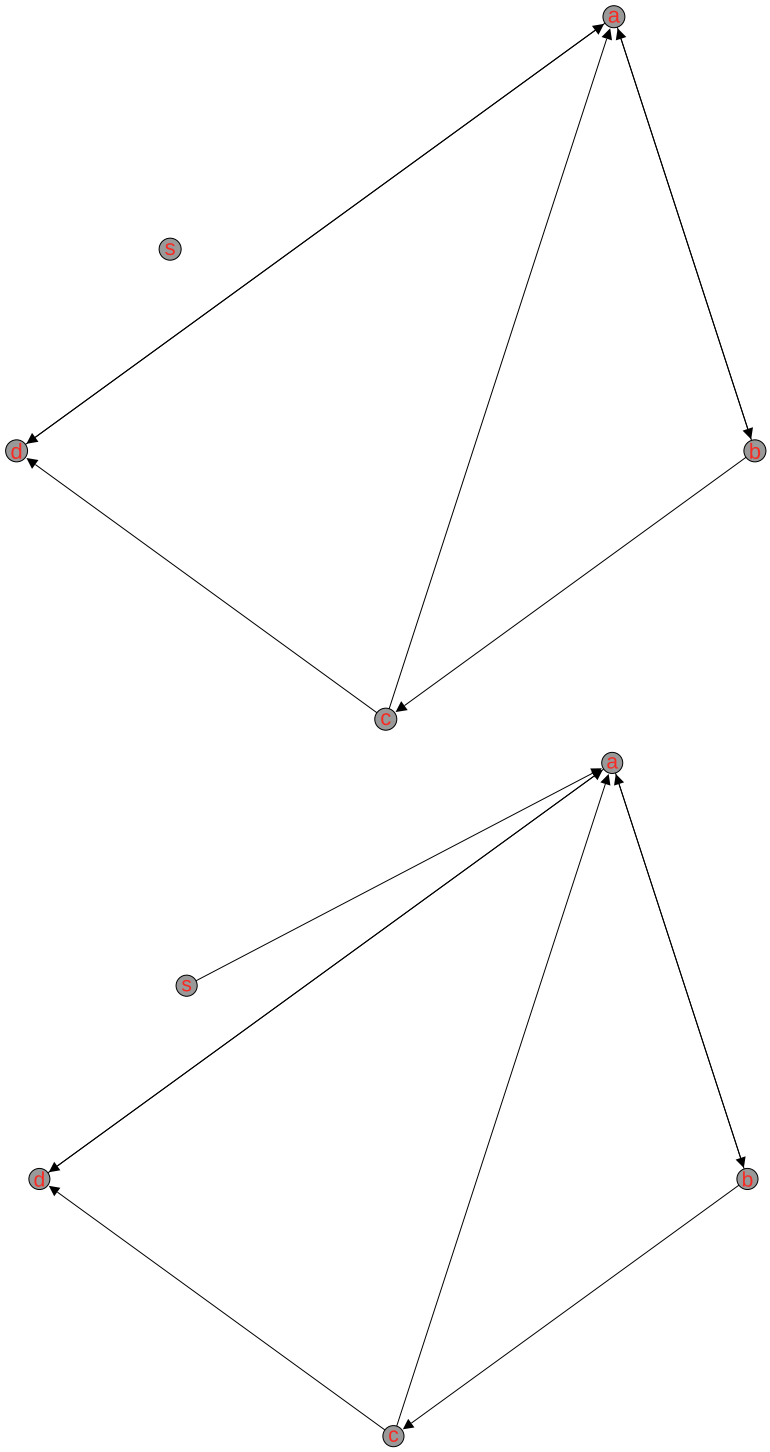


Figure 8 : Introducing vertex *s* to the network

Then by adding a new vertex t the boundary condition, before connecting it is $k(s=t,t)=0$. Then the integration constant $C(s)$ can be obtained by integrating, 2.1.1 with

$$k(s,t=s)=2\ln(s)+C(s)$$

$$C(s)=-2\ln(s)$$

$$k(s,t)=2\ln(t)-2\ln(s)$$

$$k(s,t)=2(\ln(t)-\ln(s))=2\ln\left(\frac{t}{s}\right)$$

Then through rearrangement we find

$$\frac{k(s,t)}{2}=\ln\left(\frac{t}{s}\right)$$

$$e^{\frac{k(s,t)}{2}/2}=\frac{\mathrm{t}}{\mathrm{s}}$$

$$se^{\frac{k}{2}}=t$$

$$s(k,t)=te^{-k/2} \quad (2.1.2)$$

Thus the degree distribution of a vertex $p(k,t)$ can be determined

Figure 9 :[Dorogovtsev 2002]

From figure 10 we see that $-\frac{\partial s}{\partial k}$ is a measure for the number of vertices with degree k and therefore for the degree distribution

$$p(k) \approx -\Delta s = \frac{\delta s}{\delta k}\Delta k$$

$$p(k) = A\frac{t}{2}e^{-k/2}$$

The normalisation factor A is determined using

$$\int_0^\infty p(k) = 1$$

$$1 = A\frac{t}{2}\int_0^\infty e^{-k/2}\,dk$$

$$= A\frac{t}{2}\left[-2e^{-k/2}\right] = A\frac{t}{2}[-1-2] = At$$

$$A = \frac{1}{t}$$

$$p(k) = \frac{1}{2} e^{-k/2} \quad (2.1.3)$$

So the degree distribution is of exponential form. The number of vertices with degree k decays with $e^{-k/2}$, which gives the exponential distribution.

2.2 With Preferential Attachment

Preferential attachment is the idea that a vertex will with higher probability become attached to a vertex of high degree than to vertex of low degree, [Dorogovtsev2002]. A new vertex s is introduced into the network and the probability of an edge, already attached to s, becoming attached to an existing vertex with a degree k is proportional to $k + A$, (where A is a constant with $A > 0$). Then there is a high chance this edge will attach to a vertex of high degree. We may say that the new vertex is attracted by the popularity of that vertex. In figure 11 it can be seen that vertex s has linked itself to vertex d as it has a higher degree over the other vertices.

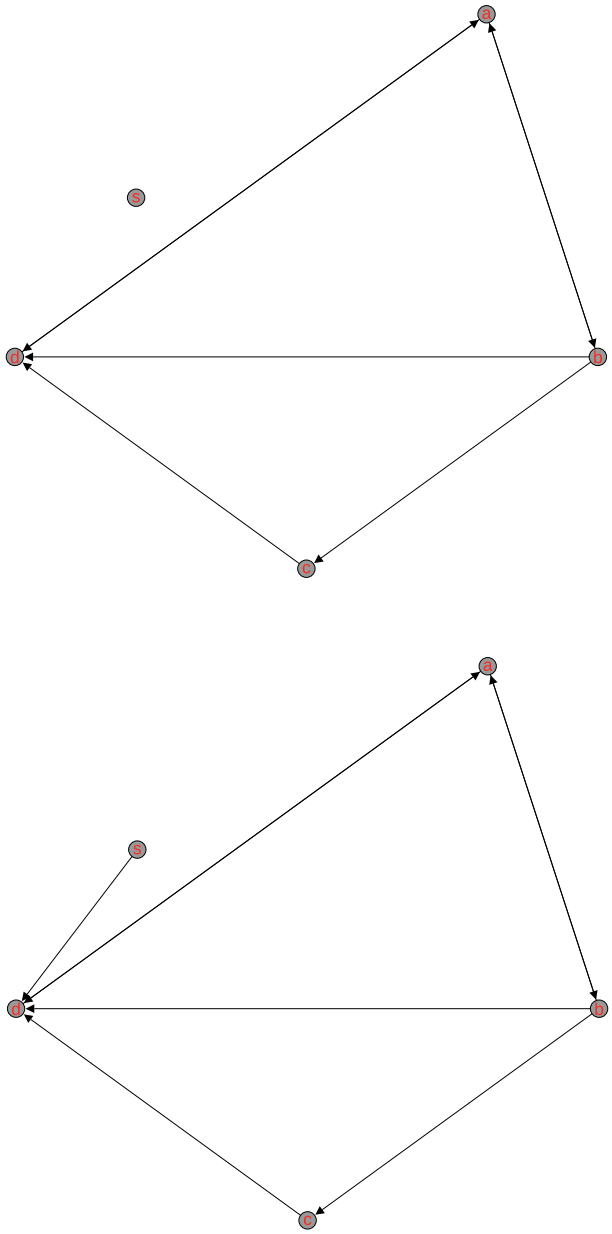


Figure 10 : Introducing a vertex

Vertex s has been introduced at time $t = s$.The degree of the vertex s at time t is $k(s,t)$. The sum of all degrees at time t must be $2t$ as the total degree is increased in each step by exactly two, corresponding to the two ends of the edge that connects the new vertex to the chosen existing vertex.

$$\sum_{u=0}^{t} k(u,t) = 2t$$

The probability for the new vertex to be attached to vertex s is therefore $2(k_s + A)$ normalised by the sum of this term for all vertices:

$$\frac{2[k(s,t)+A]}{\sum_{u=0}^{t}[k(u,t)+A]} = \frac{2[k(s,t)+A]}{[(2+A)t]} \quad (2.2.1)$$

The time evolution of the degree of the vertex introduced at time s is then determined by the following differential equation

$$\frac{\delta k(s,t)}{\delta t} = 2\frac{k(s,t)+A}{(2+A)t}$$

$$\delta k = 2\frac{k(s,t)+A}{(2+A)t}\delta t$$

$$\int \frac{1}{k+A}\delta k = \frac{2}{2+A}\int \frac{1}{t}\delta t$$

$$\ln(k+A) = \frac{2}{2+A}\ln t + C(s)$$

Using the boundary condition $k(s = t, t = 0)$, (on introducing the vertex s at time $t = s$ this vertex has initial degree $k = 0$)

$$k(s, t = s)$$

At birth time, $t=s$, of vertex s we have

$$\ln(k+A) = \frac{2}{2+A}\ln \mathrm{s} + \mathrm{C(s)}$$

which determines the integration constant

$$C(s) = \ln A - \frac{2}{2+A}\ln \mathrm{s}$$

Therefore

$$\ln(k+A) = \frac{2}{2+A}\ln \mathrm{t} + \ln \mathrm{A} - \frac{2}{2+A}\ln \mathrm{s}$$

$$\ln(k+A) = \ln(\mathrm{t}^{\frac{2}{2+\mathrm{A}}}) + \ln \mathrm{A} + \ln(\mathrm{s}^{-\frac{2}{2+\mathrm{A}}}) \qquad (2.2.2)$$

Rewriting this we finally find

$$\left(\frac{k}{A}+1\right) = \left(\frac{t}{s}\right)^{\frac{2}{2+\mathrm{A}}} \quad (2.2.3)$$

As function of t the degree $k(s,t)$ grows with a power $a = \frac{1}{1+A/2}$.

Otherwise we may rewrite (2.2.3) as

$$\frac{t}{s} = \left(\frac{k}{A}+1\right)^{1+A/2}$$

Or

$$s = t\left(\frac{k}{A}+1\right)^{-\left(1+\frac{A}{2}\right)}$$

As function of s the degree $k(s,t)$ decreases with power $-a$ given above.

For fixed time $-\frac{\partial s}{\partial k}$ is a measure for the number of vertices of degree k: and therefore for the degree distribution

$$p(k) \propto -\frac{\partial s}{\partial k} = (k+A)^{-\left(1+\frac{A}{2}\right)}$$

$$p(k) \propto \left(1+\frac{A}{2}\right)(k+A)^{-(2+\frac{A}{2})} \propto k^{-\gamma}$$

for large k where

$$\gamma = 2 + \frac{A}{2}$$

As A grows from 0 to ∞ the γ exponent changes from 2 to ∞, (it has been found that in most networks γ is between 2 and 3 as we will see later). This tells us that combining growth and preferential linking will give a power-law distribution.

As noted above the time evolution of the individual degrees is determined by the exponent $a = \frac{1}{1+A/2}$

The two exponents are therefore related by

$$a = \frac{1}{\gamma - 1}$$

$$1 + \frac{k}{A} = \left(\frac{t}{s}\right)^{a} \quad (2.2.4)$$

In this way the static scale-free behaviour is related to the time evolution of the degrees of individual vertices.

All networks following such a distribution are called scale-free networks. Examples include the author collaboration network, protein interaction and the World Wide Web, [Dorogovtsev2002]

2.3 Barabási and Albert Model

Barabási and Albert [Barabási2002] were pioneers in the exploration of such network growth behaviour and produced an early but less flexible model (it lacks a parameter A and therefore produces networks with a fixed power law exponent γ).

In this model, a new vertex s is added to the network by an edge. The other end of the edge of that vertex is attached to an existing vertex with a probability proportional to the degree.

In this model the probability that a new vertex s is connected to an existing vertex u is

$$\frac{k(s,t)}{\sum_{u=0}^{t} k(u,t)} = \frac{k(s,t)}{2t}$$

With

$$\sum_{u=0}^{t} k(u,t) = 2t \qquad (t \gg 1)$$

The equation for $k(s,t)$ becomes

$$\frac{\delta k(s,t)}{\delta t} = \frac{k(s,t)}{2t}$$

$$\delta k = \frac{k(s,t)}{2t}\delta t$$

$$\int \frac{1}{k}\delta k = \frac{1}{2}\int \frac{1}{t}\delta t$$

$$\ln(k) = \frac{1}{2}\ln t + C(s)$$

Entering the boundary condition $k(s = t, t) = 1$

$$ln1 = \frac{1}{2}lnt + C(t)$$

$$C(t) = -\frac{1}{2}lnt$$

With $C(t)$ being some constant

$$k(s, t = s)$$

$$lnk = \frac{1}{2} lns + C(s)$$

$$C(s) = lnk - \frac{1}{2} lns$$

$$lnk = \frac{1}{2} lnt - \frac{1}{2} lns$$

$$lnk = \frac{1}{2} \ln\left(\frac{t}{s}\right) = l\,n\left(\frac{t}{s}\right)^{1/2}$$

$$k(s, t) = \frac{s^{-1/2}}{t} \qquad (2.3.1)$$

In the limit of a large k this gives a power-law distribution

$P(k) \propto k^{-3} \equiv k^{-\gamma}$

Therefore, $\gamma = 3$ is the only exponent this model may predict.

2.4 Limits to the behaviour of scale-free networks

There are limiting factors to consider during the process of preferential attachment, [Newman2002]. The best way to explain these is through real world examples.

Age

The age of a vertex may cause it to not produce any more connections between itself and other vertices. The network of movie actors is a prime example of this.

As an actor gets older the number of films he will be asked to star in will decrease, or the actor may die. However the actor is still part of the network thus still contributes to the statistical properties of the network.

Money and Space

An airport contains different airlines. It is physically impossible for all airlines to belong to one airport, due the restriction of the amount of space and the amount of money that would be involved in doing this. In a network this corresponds to a vertex not producing any more connections because of the physical costs in doing so.

Information and access

In some networks there are limits to the amount of information that is available. In a webpage there may be blocking constraints on out-going links to other webpages.

These limits can be modelled as below [Dorogovtsev2002]

$$p(s,k+1) = \frac{p(s,k)f(k,a,\dots)}{\sum kf(k,a,\dots)}$$

Where $f((s,k),a,\dots)$ is a function that may dependent on its age, cost, or restricted information. As a result of this a cut-off in the power-law distribution may occur.

2.5 Summary

The majority of real world networks do not follow the classic random graph and are termed scale-free. Preferential attachment is common in networks and gives the idea that vertices that are already of high degree attract more vertices. However the limiting factors are also considered and explain how a highly connected vertex can stop having new edges connected to it.

3.0 Social roles and behavioural techniques within online discussion groups

Discussion groups are online real world social networks. there are dedicated websites to discussion forums such as Google Groups, Usenet, and Yahoo groups, [Boccaletti2006][Adamic2008]. Discussion groups can also be a part of a webpage offering a place for help or frequently asked questions. These groups are usually categorized into several topics, ages, and languages, ranging from religion and politics, to mother care and technical groups. Groups exist for any possible topic that can be imagined and the public use discussion groups for many different reasons, for help, to guide others, to voice an opinion, a community where people can be themselves.

A social role of a community online is, ‘ a combination of social, psychological, structural and behavioural attributes,” [Gleave2009] From the moment an actor logs onto a social network they display characteristics of different roles. The type of role however varies depending on the actions they take, [Fisher2006].

3.1 A Thread

Figure 11 : An example of a discussion thread

Figure 12, shows a discussion sample thread taken from the comp.ai.philosophy discussion group. There are seven actors in this thread resulting in seven vertices. From the discussion thread it can be seen that *Stephen* started the thread with *David, Kevin, Cliff, Josh, John*, and *decomyn* all replying to the thread. Without reading the thread we can determine the following, the degree of each individual, the out-degree and in-degree, the clustering coefficient, betweeness and shortest average path length.

Full names have been removed from this paragraph for data protection reasons.

Assumption: If *Cliff* is replying to *Kevin* he is also replying to *David* and *Stephen*.

An edge is directed away from the vertex indicating that the actor has replied to the thread; this is the out-degree of the actor. There is not an out edge from *Josh* (5) to *Cliff* (4), because the indentation informs that *Josh* is not replying to Cliff but replying to *Kevin* (3).

An edge is directed inwards if an actor has had their post replied to by another actor, in this example there is a directed edge from David to Stephen, this corresponds to the in-degree.

Therefore the overall degree is the sum of the in and out-degrees. In this Stephen has the highest degree count.

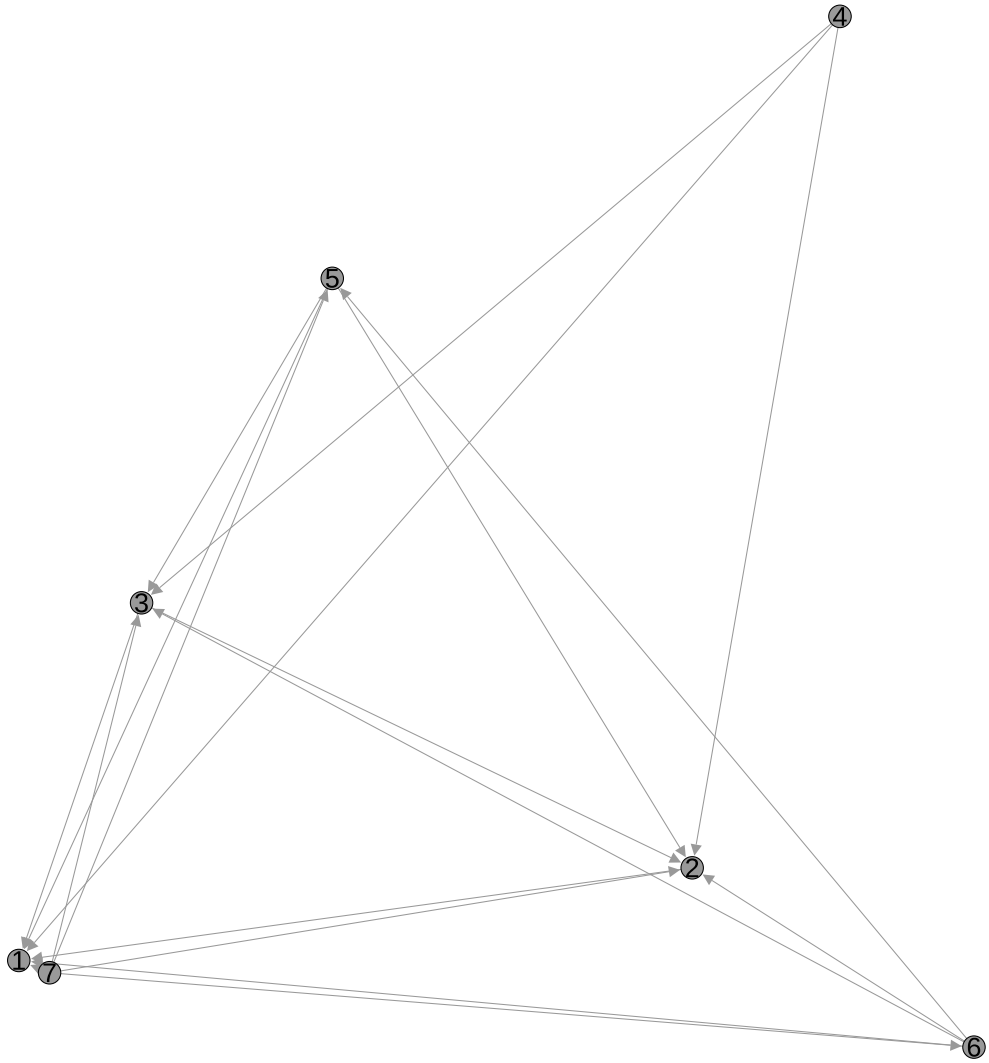


Figure 12: Network of above thread

Vertex	*In-degree* k_i	*Out-degree* k_0	*Total degree* k	*Betweeness* B	*Clustering Coefficient* C
1	*6*	*0*	*6*	*0*	*0*
2	*5*	*1*	*6*	*0*	*0*
3	*4*	*2*	*6*	*0*	*1*
4	*0*	*3*	*3*	*0*	*1*
5	*2*	*3*	*5*	*0*	*1*
6	*1*	*4*	*5*	*0*	*1*
7	*0*	*5*	*5*	*0*	*1*

Table 2 : Statistical properties of the discussion thread

	1	***2***	***3***	***4***	***5***	***6***	***7***
1	*0*	*0*	*0*	*0*	*0*	*0*	*0*
2	*1*	*0*	*0*	*0*	*0*	*0*	*0*
3	*1*	*1*	*0*	*0*	*0*	*0*	*0*
4	*1*	*1*	*1*	*0*	*0*	*0*	*0*
5	*1*	*1*	*1*	*0*	*0*	*0*	*0*
6	*1*	*1*	*1*	*0*	*1*	*0*	*0*
7	*1*	*1*	*1*	*0*	*1*	*1*	*0*

Table 3: The shortest path length from each of the vertices

Table 3 shows the way the path length is calculated for this thread. The average shortest path length is close to one as almost all vertices have a connection to each other.

The betweeness has been calculated at 0 for each of the actors this is because they are all highly connected between themselves causing a high clustering coefficient, no actor is more important than the other in keeping the network connected.

The thread count corresponds to the number of actors in the post, in this case it is the same as the number of vertices, 7, however this is not always true as an actor may post several times to the same thread.

Let us first examine the positions of the actors to provide some information on the different social properties within a thread. The position of the actor within the thread may match behavioural attributes, [Maia2008].

Behavioural attributes may match to actors who always post first (in position one), or always finishes the post (final position), or actors who always like to reply first (in second position). Although the in and out-degree are useful tools, the position within the thread provide much more information.

The reply count also provides extra information, this asks how many times a actor has replied to their own posts they have started, again some particular behaviours may rely on this property.

3.2 Social Roles

Past research has provided a wide variety of roles from the lurkers to the gender of the actor. The fascination with studying the roles online has grown over the last fifteen years with social networking sites increasing in popularity.

Below is a list of the types of roles that have been found in previous research by Welser, Gleave and Fisher [Welser2007][Welser2011] and Panzeraza, Opsahl and Carley [Panzerasa2009].

- Gregarious/popularity
- Male/female (self declared gender)
- Fans
- Trolls
- Answer people
- Discussion people / conversationalist
- Flame warriors
- Lurkers
- Debaters
- Spammers
- Question people

A number of these roles have been extensively researched. They are all roles in which actors have posted a large number of times and hence they are easier to

predict. Examining their discussion patterns may be useful to the actors and the owners of the social network site.

3.3 The Answer Role

One of the highly researched roles within the online community, and as the name suggests the main structural attribute of the answer role is the actor will be particularly inclined to answer posts over initiating posts.

Research conducted by Welser et al and Turner [Welser2007] involved analysing the content of posts and replies by actors in the discussion group, then finding common attributes among those who answered posts. This has provided a basis for the answer role for future researchers proving it is now unnecessary to read the contents of the posts. An answer person will mostly provide answers and hence will only contribute few replies to any given thread, resulting in a high out-degree and a low in-degree. The threads are expected to be short, as no real discussion will take place causing the local neighbourhood (actors connected with path length one) of the answer role producing a low clustering coefficient. They are also likely to be connected to many alters who themselves have low degree and few alters who have a high degree [Welser2007] high betweenness.

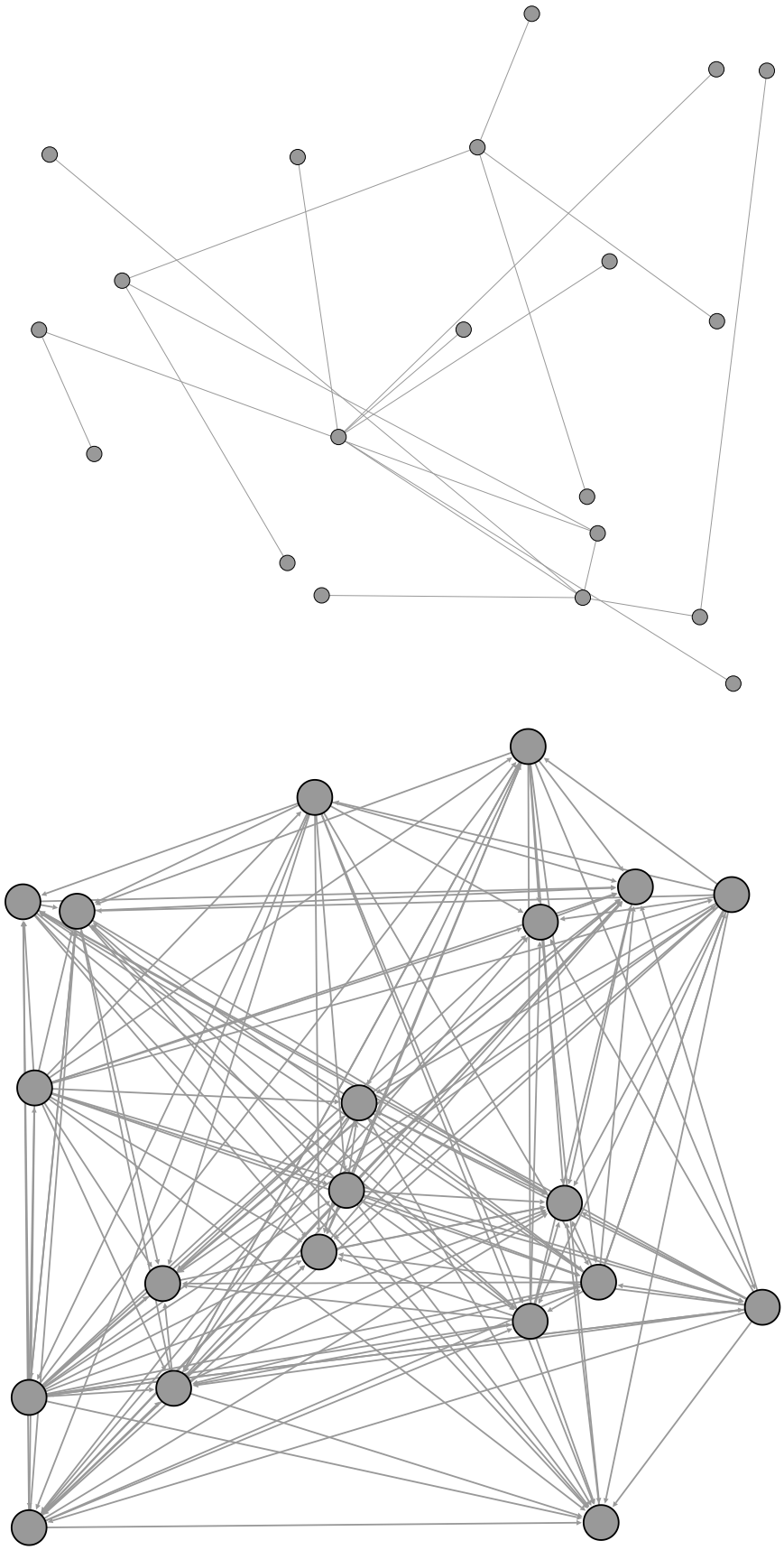


Figure 13 : (Top) Sparse network typical of answer role, (bottom) densely connected network typical of discussion role

The research in this thesis will expand on this role by introducing new measurements, the number of replies immediately to themselves (immediate reply count) and the number of replies (reply count) within the thread is expected to be minute. They are also likely to be involved in threads with few actors causing a low thread count. Time will also be investigated, if an actor has been a member of the group for a long period of time is it likely they will display the answer role? [Misolve2007]

This role however is extremely beneficiary for not only other actors but for the owners of the site. If an actor is seeking an answer they will know who to ask, and providing the answer is correct the actor may be asked a question from the same person several times, not only increasing the popularity of the actor but the popularity of the website. The more answer people involved in the website the more people will seek to ask questions. Answer actors who have been involved in the network over a long period of time are most likely to be within the larger sub groups of the network

Technical, and scientific groups are likely to have a high number of actors displaying an answer role.

3.4 The Question Role

Opposite to the answer person is the actor who asks the question. It is predominantly recognised by starting the posts and the posts are very short with only a few number of replies by another actor and possibly by themselves [Turner2005]. The egocentric graph of an actor is the graph of all connections between an actors nearest neighbour, for the question role this graph is sparse as it is mostly likely to have neighbours who display the answer role, it shows a low clustering coefficient with a high betweenness [Donath1999]

This thesis will build upon previous research by measuring the time the actor is in the group, differentiating between the out-degree and in-degree, and the thread count. It is to be expected the newer the actor the greater chance they have joined the group to ask a question. The longer the actor is involved in the group the less likely they will display the question role. The in-degree is to be high with a low out-degree. Predictions of a small thread count, similar to the answer role with the immediate reply count and reply count should be relatively low (however not as low as that for the an answer role).

It is impossible to have a group with actors displaying an answer role and not a question role. Hence those groups that are likely to have high number of question roles should also display a high number of answer roles, but not necessarily the other way round. Thus it is expected, technical and scientific groups would have a high question to answer ratio. This will also be measured further in the report.

Aside from the obvious technical and scientific groups, one may find a large number of actors displaying the question role in religious, political and language discussion groups.

3.5 The Discussion Role

A discussion role is one who seeks to have a conversation with other actors within the particular topic of the group [Turner2005]. Known by other researchers as a conversationalist.

It has been found a discussion role will not only initiate threads but will also respond, having a high number of replies to posts initiated by either themselves or others, causing a high out-degree. This out-degree will be similar to in-degree as the main aim of the discussion is to expect many replies from a post. Majority of the discussion role neighbours will also be discussion roles resulting in a densely connected local neighbourhood [Turner2005] with a relatively large number of connections [Welser2007]. This will cause the clustering coefficient to be high with a low betweenness. The lengths of the post within the threads they initiate or belong to are long [Welser2007].

The current thesis aims to improve the prediction of the discussion role through further investigation into the immediate reply count, the thread count, and the average posts per month. It is anticipated that along with a high in and out-degree

the thread count will also be high, due to the idea a discussion role not only seeks a topic to discuss but also initiates one. The immediate reply count may also prove a useful measurement as it is expected threads will be long and one can assume there may be a limit to the number of characters used in any given post and hence may need to immediately reply to themselves to continue the post. The thread count will be lengthy along with the number of posts per threads, causing the average posts per month to be high in relation to an answer role and a question role. Time is not expected to be a factor as an actor displaying the discussion role may display the properties almost immediately.

The benefits of having a discussion role within an online social network group is that it will encourage the popularity of the group, bringing in new actors. Groups that are likely to display actors with the attribute of the discussion role are religious and philosophical groups.

3.6 Spammers

Spammers within a discussion group are actors who send unwanted posts, often advertising or junk. Attributes displayed by these are high out-degree with extremely low in-degrees. As interactions between actors rarely happen they will have a scarce local neighbourhood with a low clustering coefficient and high betweenness [Turner2005]. For the overall structure of this group the spammer is

likely to be found in isolated small sub-groups with a small amount in the central hub of the group.

Research into this role could be expanded through analysis of the time an actor has been with the group, number of posts started and given replies to compared to the number of incoming posts and the use of links within the post referring an actor to another site.

Spammers are an unwanted nuisance in a discussion group. A group with a high number of spammers will deter any other actors joining the group. By detecting spammers early, owners or moderators of the site can block any further posts. A spam free discussion group is highly desirable. However spammers can be found in any discussion group.

The roles above are the main ones that can be found in groups and that can determine the type of group. The roles below are a brief breakdown of other smaller roles that can also be found.

Gregarious and Popular

The gregarious and popular is one of the roles that have been investigated by Panzarasa [Panzarasa2009]. This role shows a sociable relationship with a high in and out-degree.

Male and Female,

Panzarasa also demonstrated it was possible to determine differences in posting behaviour of a male and a female actors. It is believed that men initiate and reply to posts more than women thus causing men to have a higher out-degree than women and women to have a higher in-degree than men.

Fans

Welser et al investigated the role of a fan, through accessing the posts and reading them. They systematically express appreciation or affiliation and hence this would prove difficult to predict without reading the posts.

Trolls

The main aim of a troll is to create useless discussions. Identification of these posts would need usually to be human read [Welser2007]. These individuals have a high number of posts in they have initiated, resulting in a high out-degree.

Flame warriors

Flame warriors are similar to discussion roles in that they are involved in long threads, however their one aim is to cause disruption within the discussion through endless unfriendly discussion on often minor issues, see [Welser2007].

3.7 Position of Posts

The position of the actors post within the thread gives more insight on the actor's behaviour. All positions within the thread are of interest as the following explains.

The First Post

This first post is the actor who is initiating the conversation, the one who asks a question or merely wishes to discuss a topic. It is expected that actors with a high amount of first posts when compared to other positions within the network are categorized by the question role. This does not mean that all actors who post first are put into the question category. The discussion role will also have a proportion of their posts in the first position; this will usually be similar to posts within in other categories. Spammers could also fall in to this category with a high number of first posts.

The Middle Post

Assuming a post has more than one reply, the middle posts are the in-between posts. One would not expect an answer role or a question role to be displayed here. Discussion roles will be highly dominant here.

The Last Post

Assuming the post has at least one reply, the last post is the post that finalises the thread. One would expect the actor with a small number of last threads to display a question role. An answer role would display a relatively large number of last posts.

A discussion role may also display a proportionate number of last posts similar to posts they have started.

4.0 Data

Initially this research will explore the behavioural roles of two groups, comp.ai.philosophy and comp.text.tex. Both groups can be found on Google groups website http://groups.google.com/. These groups were chosen due to the large number of actors and the high number of posts per month and due to the fact that they may display differences resulting from the different topics. .

The research uses quantitative analysis to investigate the behavioural attributes of online discussion groups within Google groups. This will be processed through the following;

- The structure of the group and individuals through properties previously measured.
- Applying new measurements and selecting high actors in each group to predict their role within the group
- Investigate the time evolution of the group

For both groups the following mathematical properties will be investigated to determine if the initial prediction of both groups are correct

For individual actors within the group

- The Degree Distribution – The in-degree, out-degree, reply degree and immediate reply degree.
- The clustering coefficient
- Density (how close the network is to complete)

- Thread count
- Post Length
- Time
- Average posts per thread
- Posts positions

For the whole group

- Top five actors
- Network Size – the number of actors, number of connections, diameter, average shortest path lengths, and presence of sub-graphs
- Average thread count
- Time span investigated
- Average posts per month
- Post's positions

Comp.ai.philosphy **-** As the name suggest discussions within this group cover the topic of philosophical aspects of Artificial Intelligence. The group initially started back in 1990 and is still very heavily active to this present day. This group is expected to be predominantly filled with discussion roles, consisting of long threads and high thread counts, a large and densely connected network.

Comp.text.tex **–** A technical group with topics concerned with discussion about the TeX and LaTeX systems and macros. The group still available today also dates back to 1990. Predictions for this group include high number of answer and question

actors, a large sparsely connected network with low clustering coefficients and high betweenness of a number of active individuals.

4.1 Sample Groups

A small sample of both groups is initially sampled and analysed. The sample was taken by randomly selecting a months post of each group. Each post is read thoroughly and then categorised. The posts will be categorised as follows

- Answer Post
- Discussion Post
- Question Post
- Spammer
- Announcements
- Statements

Answer Post: If a post answers a question from a previous post. The actor would not give the impression that they are expecting a reply.

Discussion Post: Posts are categorised as discussion when the actor does not ask for help or provide answers. The post is simply a comment that wishes to gain response from other actors, creating an informal debate.

Statement: Posts containing only one single sentence or a comment that does not wish to cause a reply.

Question: When a posts main objective is to ask for help on a problem or some form of guidance or support.

Announcement: Posts that contains updates of software or hardware.

Spam: Posts have no relevance to the group and cause nuisance by advertising or selling items.

After reading and categorising the posts it was noticed some posts displayed more than one category. As can be seen below categories such as ans/que and que/dis overlap. Perhaps an actor answered a question and felt they should elaborate on it more causing the need for discussion. Posts were given the following codes at end of each post.

Sta – Statement

Ans – Answer

Dis – Discussion

Que – Question

Que/Dis – Question and discussion

Que/Ans – Question and answer

Ano – Announcement

Ans/Dis – Answer and discussion

Briefly comparing the two groups results in figure 15 below one can observe distinct differences.

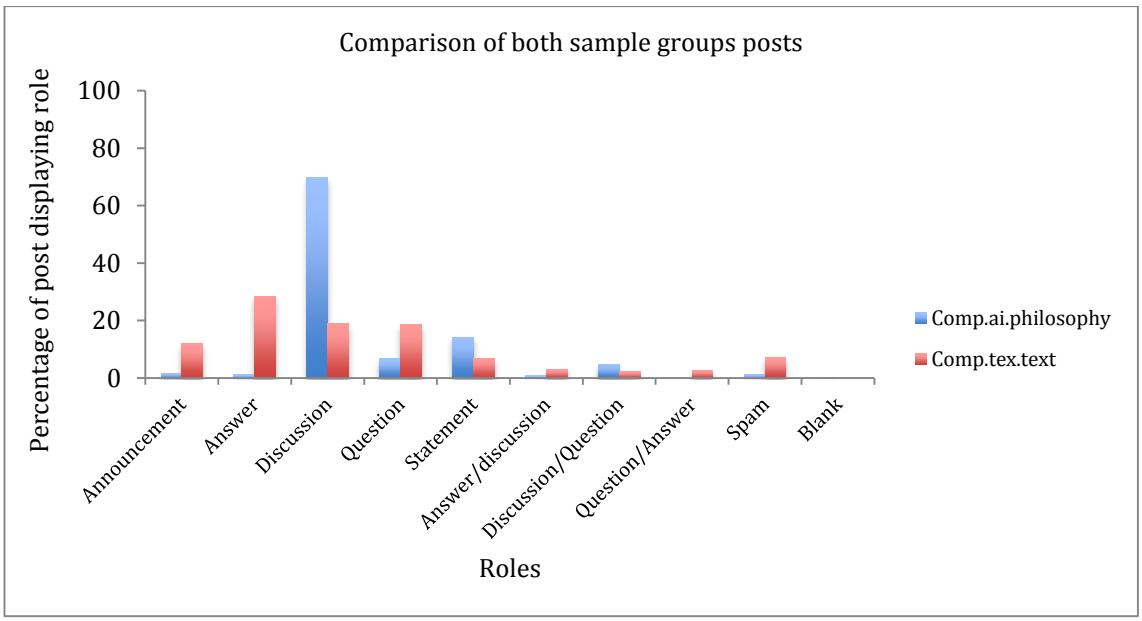


Figure 14 : Sample groups histogram

Comp.tex.text: 521 posts

Comp.ai.philosophy: 663 posts

The major difference is within the Comp.ai.philosophy which has a high percentage of its posts within this sample display in the discussion property. Comp.tex.text has more sporadic categories with the highest number of posts display in the answer property followed closely by the question and discussion posts.

Looking at each group individually we treat both samples as a complete network. By delving further into each type of category through individual actors, the statistical properties of the sample networks and the position of the posts.

Gathering all this data against the categories will help to predict the actors and the posts for the remaining global network.

4.1.1 Comp.ai.philosophy Sample Group

As previously seen the comp.ai.philosophy sample group already displays a vast amount of posts that are within the discussion category. Examining the statistical properties of this sample network;

Number of Vertices, V	108
Number of Edges, E	1229
Average Degree, $\bar{k}$	11.38
Average Weighted Degree, $\overline{k_w}$	154.92
Diameter, D	5
Graph Density, G_D	0.106
Clustering coefficient, $\bar{C}$	0.59
Average Shortest Path Length, $\bar{l}$	2.083
Betweeness, $\bar{b}$	0.007

Table 4 : AI sample group properties

All the above results point to a densely populated network, the high average degree and weighted degree, high average clustering coefficient, small average path length and diameter, and small betweeness.

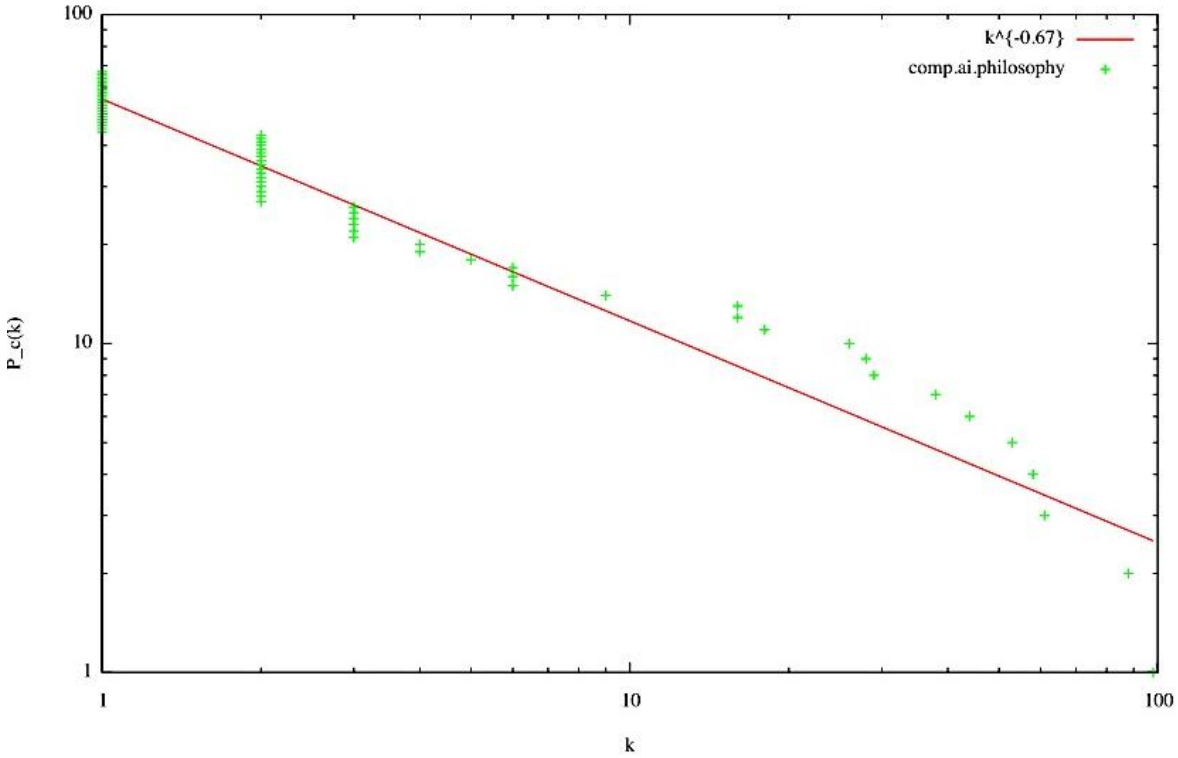


Figure 15 : Cumulative degree distribution for Comp.ai.philosophy

The plot shows the discrete inverse cumulative degree distribution on a log-log scale, fitted with a power-law line of best fit. Due to its discrete nature this cumulative distribution displays a 'staircase' behaviour. In the graph each count for the same degree is indicated by a separate tick-mark creating a small column for each degree. From this cumulative plot the exponent of the power-law can be derived, using the following equation.

$$\gamma = 1 + \gamma_{cum}$$

$$\gamma_{cum} = 0.67$$

$$\therefore \gamma = 1.67$$

Most real world networks will have $2 \leq \gamma \leq 3$ and this falls just under this boundary.

The high overall degree coupled with a high weighted degree is the cause for such a low average shortest path length and diameter. There are 64% of actors who have greater than the average degree, and 28% of the actors have higher than the weighted degree. This confirms there are more discussions between any two actors. With an average shortest path length of only 2.083 informs that out of the 108 actors any one actor can reach another actor through 1 person. The diameter is the maximum shortest path length, taking two actors a, b there is a maximum of 4 other actors between them, a result of a highly connected network.

The global positions of the posts of the network, three positions the posts can take, first, last or middle position. Figure 17 shows the position of the posts for this sample group.

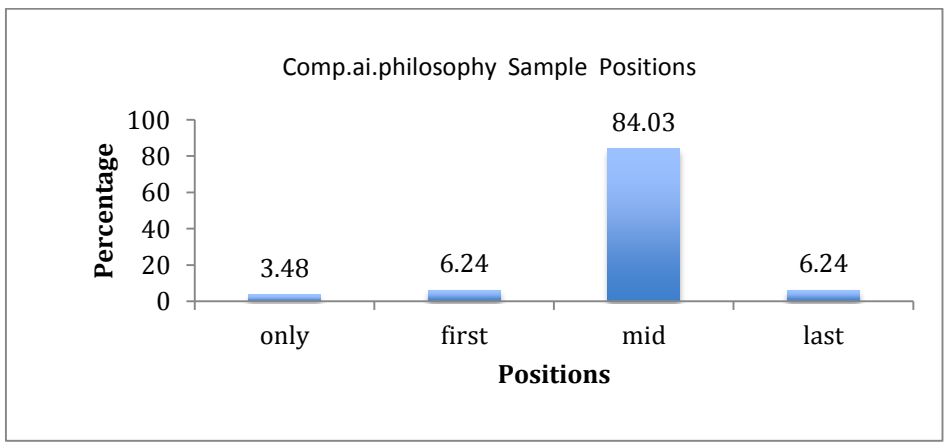


Figure 16 : Post positions for Comp.ai.philosophy

Overall the group have 84% of its posts in the middle position suggesting there is a great amount of conversation between actors in the same thread over actors initiating new threads and finishing threads. There is also a small amount of posts that are threads in the only position, which relates to posts that are idle, meaning threads containing only one post. This could correspond to the small number of spam posts.

.

The results from the overall group suggest for a group to be predominantly filled with discussion role it should have the following properties.

- High average degree, both weighted and non weighted
- High number of actors that have greater than average degree and low amount of actors that have greater than the average weighted degree.
- High clustering coefficient
- Low average shortest path length
- Large amount of posts in the middle position
- Small average betweenness

Further investigation into individuals who display not only the discussion role but also the question and answer role is needed. This will help create a more solid structure of the properties of the individuals further solidifying the group properties.

Find below the top three actors who display the discussion role All three actors give a large amount of contribution to the posts with each having a large percentage of the posts in the discussion role

- WC01
- TT01
- CA01

	WC01	**TT01**	**CA01**
No. of posts	81	31	70
Average Degree, $\bar{k}$	110	63	81
Average in-degree, $\bar{k}_i$	62	31	54
Average out-degree, $\bar{k}_o$	48	32	27
Average Weighted Degree, $\bar{k}_w$	4575	2138	3591
Average Weighted In-degree, $\bar{k}_{w,i}$	2040	948	1394
Average Weighted Out-degree, $\bar{k}_{w,o}$	2535	1190	2197
Clustering Coefficient, $\bar{c}$	0.19	0.31	0.22
Betweeness, $\bar{b}$	0.099	0.052	0.062
Closeness	1.44	1.65	1.86

Table 5 : Top 3 actors to display discussion role

In table 5 all three of these actors follow similar properties, the in-degree is higher than the out-degree (for both weighted and non-weighted) , and similar closeness values. The clustering coefficient is the only property that does not give a similar value for all three actors, however this would still be classed as relatively small.

Continuing with the posts positions as can be seen below all three actors have a vast amount of their posts in the middle position.

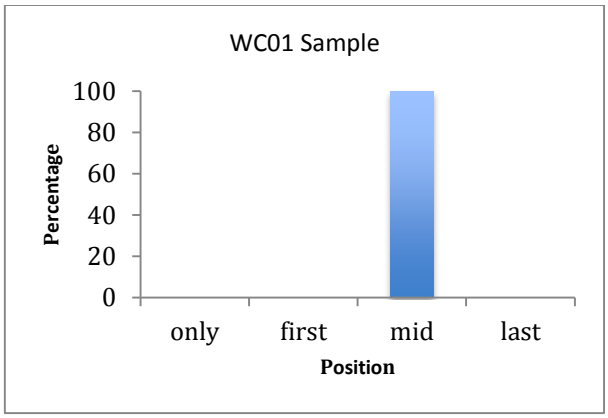


Figure 17 : WC01 Sample

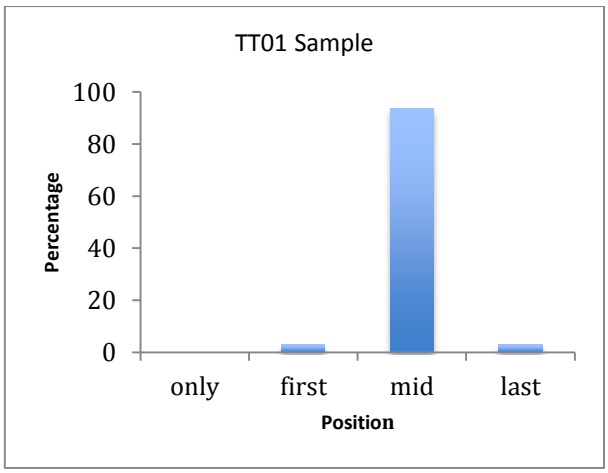


Figure 18 : TT01 Sample

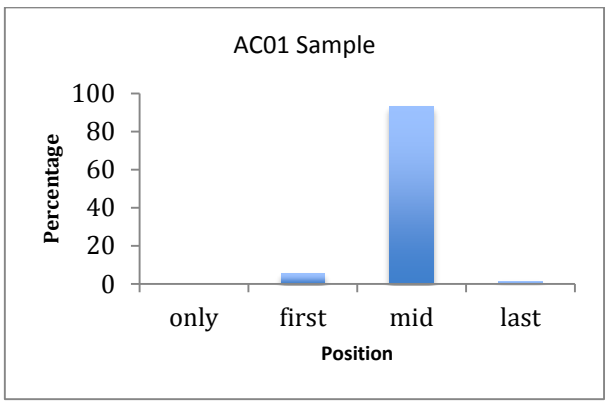


Figure 19 : AC01 Sample

Comparison should then be made between an actor with the discussion posts and actors who have majority of their posts in the question or answer post.

	Question	Answer
	DR01	AT01
No. of posts	11 with 54% as question	3 with 67% as answer
Average Degree, $\bar{k}$	80	7
Average In-degree, $\overline{k_\iota}$	29	3
Average Out-degree, $\overline{k_o}$	51	4
Average Weighted Degree, $\overline{k_w}$	573	11
Average Weighted In-degree, $\overline{k_{w,\iota}}$	262	5
Average Weighted Out-degree, $\overline{k_{w,o}}$	311	6
Clustering Coefficient, $\bar{c}$	0.044	0.081
Betweeness, $\bar{b}$	0.01	0.0001
Closeness	0.20	0.33
Average post position	64% in the middle	67% in middle

Table 6 : Question and answer role's

Posts positions for these two actors are given in table 6 above. The statistical properties of both actors are slightly different to these actors who had discussion posts, in the fact that they both have less incoming posts than out going posts. Unfortunately one cannot determine further results from the positions of these actors as both only posted a small number of times. Thus looking into the Comp.tex.text group's results that have more question an answer actors may provide further information.

4.1.2 Comp.tex.text sample group

As a technical discussion group it was expected that this group is to have more question and answer actors than discussion posts. Thus true to form after analysing each post the highest number of posts was in fact answer posts. However the quantity of discussion posts and question posts are similar. Properties of this sample group are given below.

Number of Vertices, V	142
Number of Edges, E	759
Average Degree, $\bar{k}$	5.754
Average Weighted Degree, $\overline{k_w}$	47.127
Diameter, D	6
Graph Density, G_D	0.038
Clustering coefficient, $\bar{C}$	0.151
Average Shortest Path Length, $\bar{l}$	2.785
Betweeness, $\bar{b}$	0.006

Table 7 : Properties of sample group for Tex group

All above results differ strongly from the results of the sample group for comp.ai.philosophy. There is a lower number of connections between a higher number of actors which gives a low degree and weighted degree. The clustering coefficient is also low due to few connections between actors. The diameter is much larger, suggesting this group is much more sparsely connected. The only common property the both sample groups display is a small shortest path length.

The graph density and betweenness is also extremely small, where a density or betweenness value nearer to one suggests the graph has all possible edges connected.

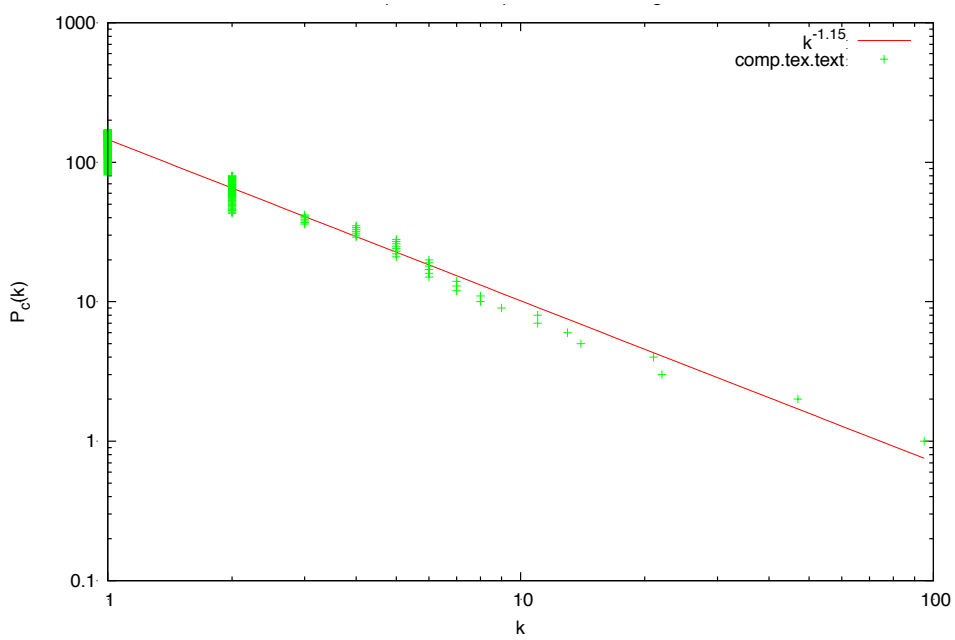


Figure 20 : Cumulative degree distribution for Comp.tex.text sample

Like the AI group this cumulative degree distribution is shown in figure 21 also shows a column degree. In this distribution $\gamma = 2.15$, which is expected for most real world networks.

The low degree and low weight degree could be the cause for a high diameter of only 6 shortest path lengths. 49% of the actors have a degree higher than the average degree and 28% of the actors have a higher weighted degree than the average weighted degree.

The average shortest path length for this sample group, although still relatively small, is higher than for the comp.ai.philosophy sample group. For any two actors there is on average one to two actors between them.

Examining at the global posting positions for the group, it can be seen in figure 22 that the majority of actors are in the middle position. Although this is unexpected for this sample group and is also limited by the number of thread, when compared to the comp.ai.philosophy group there is a much more wide spread of posts positions.

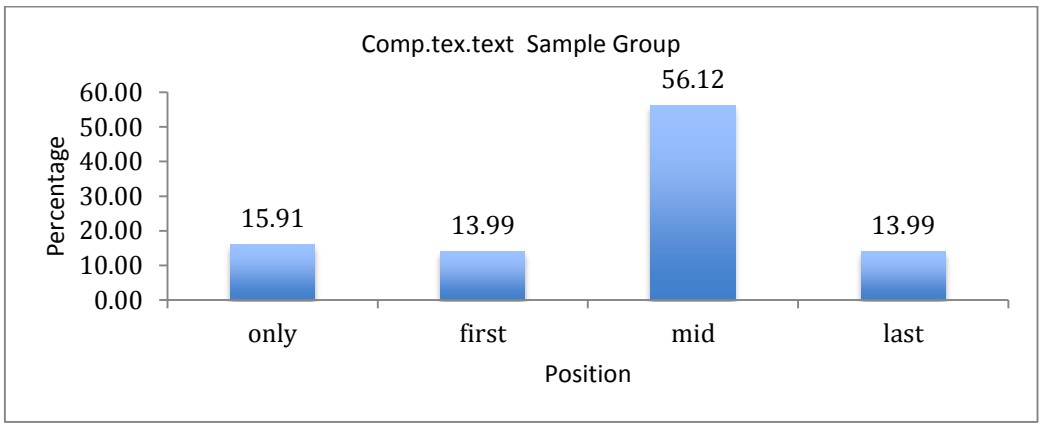


Figure 21 : Post's positions for Comp.tex.text sample group

One can conclude from the information given above that for a group that is expected to have the majority of its actors asking questions and answering questions will follow the properties below.

- Low average degree for weighted and non-weighted
- Low clustering coefficient

- Low average shortest path length
- Wider spread of posts positions
- Low graph density

On examination of individuals who have the majority of their posts as questions or answers and by comparison to the discussion actor's guidance to the properties of each type of actor is determined. In table 8 the properties of each type of actor are provided.

	Question	Discussion	Answer
	BH01	SS01	GE01
No. of posts	10	11	6
Avergae Degree, $\bar{k}$	19	22	19
Average In-degree,$\overline{k_\iota}$	10	11	10
Average Out-degree, $\overline{k_o}$	9	11	9
Average Weighted Degree, $\overline{k_w}$	135	470	30
Average Weighted In-degree, $\overline{k_{w,\iota}}$	76	194	15
Average Weighted Out-degree, $\overline{k_{w,o}}$	59	276	15
Closeness	2.33	2.36	2.86
Betweeness, $\bar{b}$	0.0018	0.000519	0.0168
Clustering Coeffciient, $\bar{c}$	0.65	0.87	0.21
Average post position	60% middle	92% middle	100% middle

Table 8 : Question, discussion and answer role's

Each of these actors have majority of their posts in the middle position that does not agree with what was expected. Looking at the degrees of the actors for the

non-weighted degree the in-degree is greater than the out-degree and this follows suit for the weighted degree for the question actor. This is not the case for the weighted degree for the answer actor with equal in-degree and out-degree. Although the posts positions do not give a great deal of information other actors that displayed each of the roles do not have enough posts to draw conclusions from. Thus further investigation to into the whole of the group would help solidify question and answer roles within the network.

Further investigation is needed to confirm if these two groups have predominantly question and answer actors or discussion actors.

New measures introduced include the reply count and the immediate reply count. The reply counts the number of times an actor has replied to a post that they have started or already replied to. The immediate reply count is the number of times an actor replies immediately after they have already posted. It is expected that the group is to have a higher percentage of reply count and immediate reply count than that of the Tex group, as there are more discussion actors.

Table 9 shows the results for the reply count and the immediate reply count. It is clear that this is not the same as the expectations with the Tex sample group having a greater reply count and immediate reply count.

	Reply Count (percentage of posts)	Immediate Reply Count (percentage of posts)	Total Edges
Comp.ai.philosophy	3.58	0.24	1229
Comp.tex.text	8.03	2.89	759

Table 9 : Reply count and immediate reply count for both groups

4.2 Summary

The Artificial Intelligence group has a majority of posts in the middle position, similar in-degree and out-degree, and high clustering coefficient. Individuals within this group also held these expected results. If the entire group displays these properties it would confirm the group would be predominantly discussion. The only property that does not follow this trend is that of the reply count and immediate reply count. As this was only a small sample of the entire group it may be that this result is higher in the main results.

The Tex group did correspond to some of the expected results including the clustering coefficient, diameter and graph density. The biggest difference that was unexpected is the posts positions, with majority of the actors in middle position. A group with the majority of its post's in the middle position would correspond to the discussion role, however although this group had high middle position it did have a greater number of posts in first last and idle positions when compared to the Artificial Intelligence group. In the main results, looking at these positions and comparing it to those in the Artificial Intelligence group will help to establish the role type.

5.0 MAIN RESULTS

After the sample results confirmed certain properties of the two networks, the following hypothesis is formed from the sample results of the two discussion groups

Hypothesis: By studying the statistical properties and posts positions of two discussion groups, expectations of each group include:

- Comp.ai.philosophy group will consist of mainly discussion role's, with the following properties:
    - Comparing the two groups, Comp.ai.philosophy is to have a greater average degree and average weighted degree than the Comp.tex.text group.
    - Low betweenness
    - High clustering coefficient
    - Small diameter
    - Small average shortest path length
    - Large percentage of posts in the middle position
- Comp.tex.text group is expected to consist of a mostly the answer role followed closely by question and discussion roles, this is expected to produce the following properties:

- Higher number of vertices and edges than the Comp.ai.philosophy group
- A small average shortest path length and diameter because of the size of the network, however this value is expected to be higher than the AI group.
- Low global clustering coefficient
- Low global betweeness, however this value is expected to be lower than that of the AI group.
- Large percentage of posts in the last position followed by the start and middle

Table 10, shows the properties of both groups, and initially differences can be seen when compared to one another. Tex group is the larger group with a greater number of edges and vertices than the AI group. Therefore comp.ai.philosophy has a greater average degree, weighted degree, density and clustering coefficient. The comp.tex.text group has a greater number of vertices, and edges. Each of these differences will be explored further below and compared to the results of the sample groups and the random graph, (Albert Barabási scale free graph) [Molloy1995].

	Comp.ai.philosophy	Comp.text.tex
Number of Vertices	3783	14264
Number of Edges	73907	149654
Avg. Degree	19.537	10.492
Max in-degree	1033	2558
Max out-degree	1205	3804
Avg. Weighted Degree	1664.49	65.099
Max Weighted In-Degree	1049690	39841
Max Weighted Out-Degree	1027464	45733
Network Diameter	9	9
Graph Density	0.005	0.001
Avg. Clustering Coefficient	0.651	0.472
Avg. Shortest Path Length	2.901	2.884

Table 10 : statistical properties of both groups

5.1 Degree and Degree Distribution

The average degree and average weighted degree for the AI group is greater when comparing the Tex group and has a larger number of vertices and edges. This suggests there are increased multiple interactions between any two actors in the AI group than in the Tex group. This result mimics that of the sample groups.

For the Tex group, there are 30% of its actors who have greater than or equal to the average degree which is less than the AI group with 46%. For the weighted degree Tex group has 14% of its actor greater or equal to the average where as the AI has only 5%. This shows the degree of actors for the Tex group many vertices with small degree, where as the AI group would have a greater range.

The discrete inverse degree distribution can be seen below in figure 22, of both groups. Clearly visible in the AI group a greater number of actors with any given degree than the Tex group. The decay of the graph is much faster for the Tex group which confirms the above comment that there are many vertices with small degree.

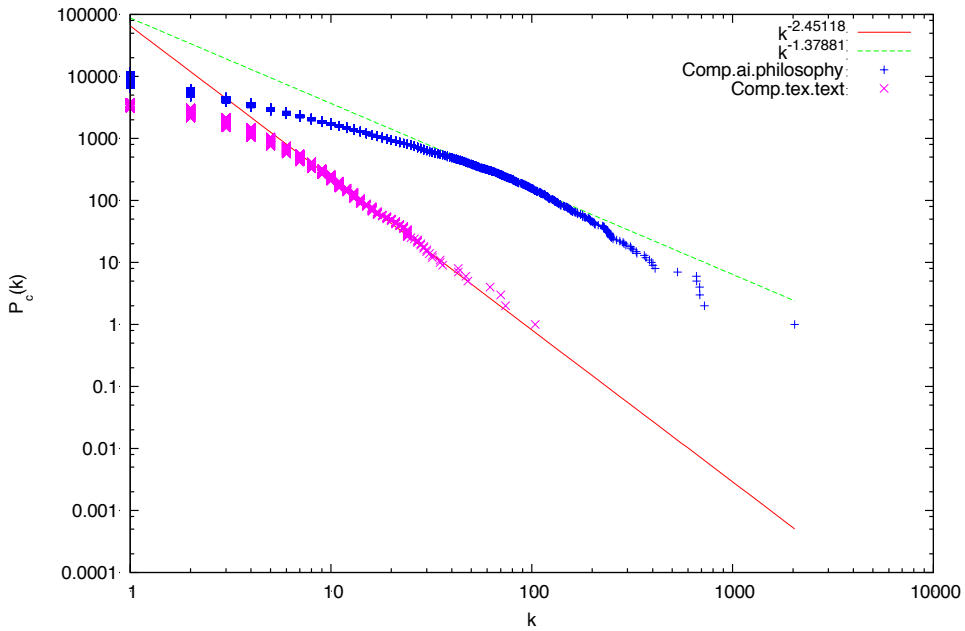


Figure 22 : The inverse cumulative degree distribution for both groups

The weighted degree is important for the analysis of the roles of these groups and the individuals. Take two actors, *a*, and *b*, if *a*, comments on a post that *b* has started there will be one edge between the two vertices in the graph. If *a*, then posts three more times onto posts that *b* has initiated on the graph this will still be displayed as one edge however it will have a weight of four, as it will represent four posts. Results show that the weighted degree is always higher than the non-weighted degree and rightly so, therefore comparison between the degrees is needed. In the AI group, when the weighted degree is divided by the non-weighted

degree it results in 85 (rounded), this means that on average for every one edge displayed on there are 85 connections (not taking into account the direction and hence will represent both in and out-degrees). For the Tex group this is significantly smaller with on average one edge representing only 6 connections. Again confirming that the AI group has more interaction between same two actors than the Tex group.

Looking further into this, comparison between the in and out-degree of the two groups is explored. Expectations are that the AI group should have a similar amount of in and out-degrees, where as the Tex group should have a higher number of out-degree than in-degree as it is expected to consist predominantly of an question and answer type of group. Initially looking at the maximum for the in and out-degrees and both weighted and non-weighted, it is certainly true for the Tex group with the out-degree at 1.49 times greater than that of the in-degree for non weighted and 1.15 times greater for the weighted maximum. For the AI group the out-degree for non weighted maximum is 1.14 times greater than that of the maximum in-degree. The weighted results for the in-degree is 1.02 times greater than the out-degree.

A discussion role should have a similar number of in and out-degrees and should be fairly close to the $x = y$ line. A question role is expected to have a high in-degree when compared to that of the out-degree and hence actors should be around the bottom right corner of the graph, provided in figure 24. Answer role's should be around the top left corner of the graph as it is expected to have high out-

degree and low in-degree. Then spammers are expected to be in the left side of the graph.

The graphs displaying the in and out-degrees (weighted and non weighted) have a large amount of data points with a positive correlation, many of which do show the in and out-degree to be proportionate to each other. Both groups do also show the actors displaying all four roles. Although it would seem most of the data points do show the discussion property. In the Tex group there are a few actors that are noticeable show properties of the question and answer role. However the significant difference between the two groups is that the AI group has more actors who have a high in-degree and out-degree.

Top five actors of highest degree are shown in table 11 below. For the AI group the top actors are different in both degrees. However for the Tex group they have the same top actors but in a different order. This shows that the top actors of the AI group have a greater amount of posts in one thread that the Tex group.

Comp.ai.philosophy		Comp.tex.text	
Weighted	Non-weighted	Weighted	Non-weighted
ZL01	ZL01	FR01	FR01
OT01	WC01	KD01	FU01
KJ01	KW01	ML01	ML01
WA01	SG01	FU01	AD01
CA02	LD01	AD01	.Á Í Å8ÒÅ Í ÏÖÅÄ

Table 11 : Top actors for both groups

Full names have been removed from this paragraph for data protection reasons.

New measures introduced that are closely linked to the degree are the reply count and the immediate reply count. The reply count is the number of times a actor would reply to a thread they had either started or already posted on. The immediate reply count is the number of times a actor replies immediately to a post they have just posted. All results can be seen in table 12, which is represented by a percentage of total posts for weighted and non-weighted.

Reply Count	Non-weighted		Weighted	
	%	Total posts	%	Total posts
Comp.ai.philosophy	1.81	73907	14.03	6296779
Comp.tex.text	4.35	149656	15.84	927857

Immediate Reply count	Non-weighted		Weighted	
	%	Total posts	%	Total posts
Comp.ai.philosophy	0.431	73907	0.039	6296779
Comp.tex.text	2.78	149656	2.029	927857

Table 12 : Reply count and immediate reply count

It was expected that the AI group would have a higher reply count than that of the Tex group. Although this is not the case in terms of percentage of the total posts for the groups, if one does not take into account the total posts and simply looks at the total number of reply for the weighted the AI group has 883249 posts compared to the Tex group with only 146957. This is also seen in the immediate reply count where the Tex group has the greater values.

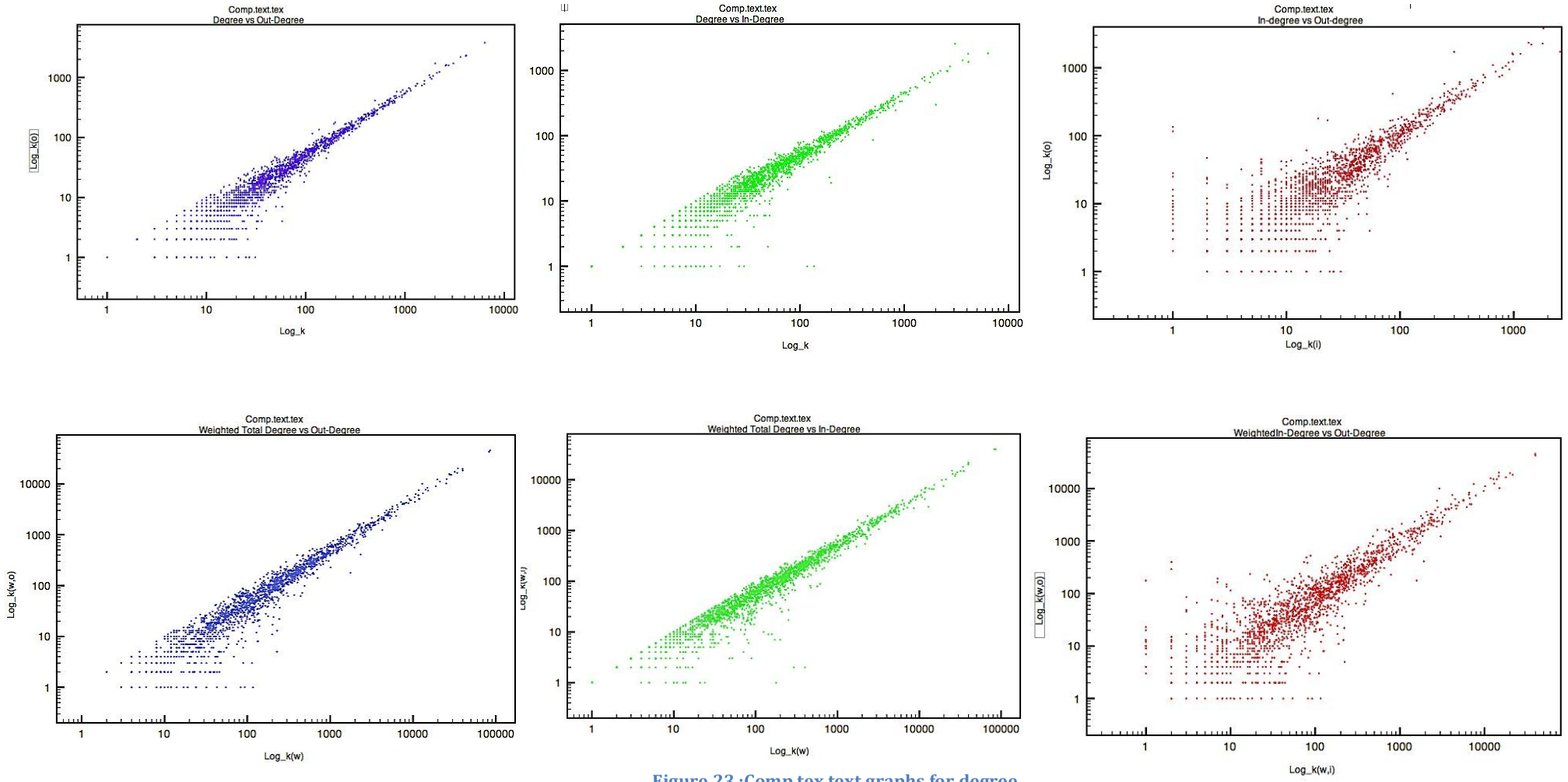
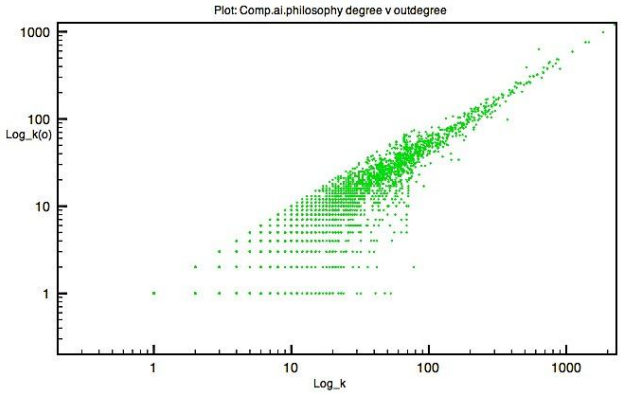
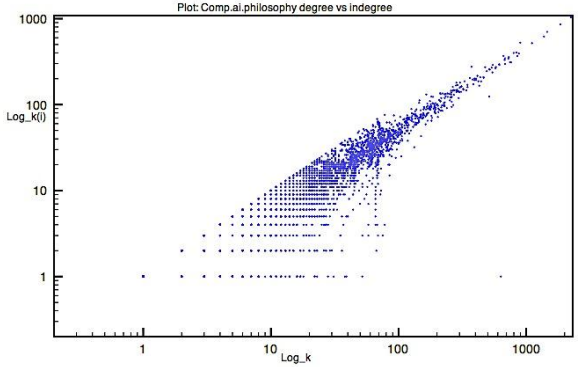
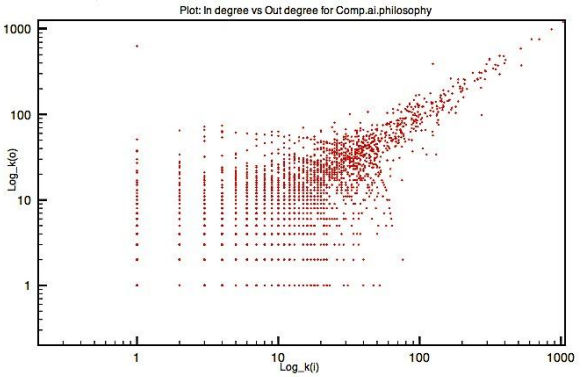
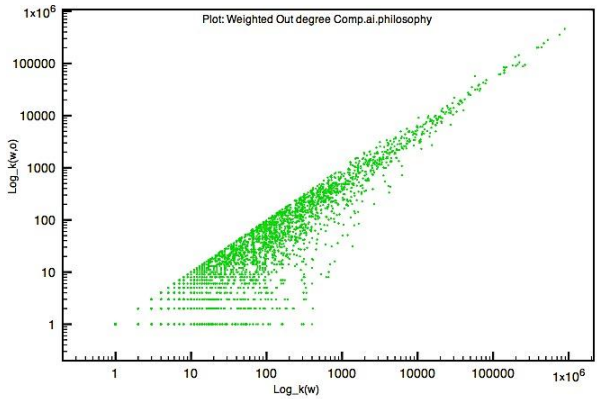


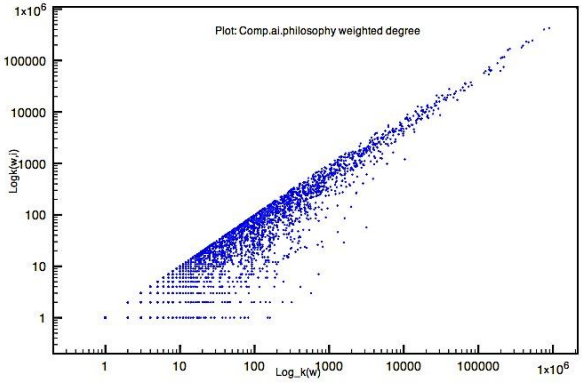
Figure 23 :Comp.tex.text graphs for degree











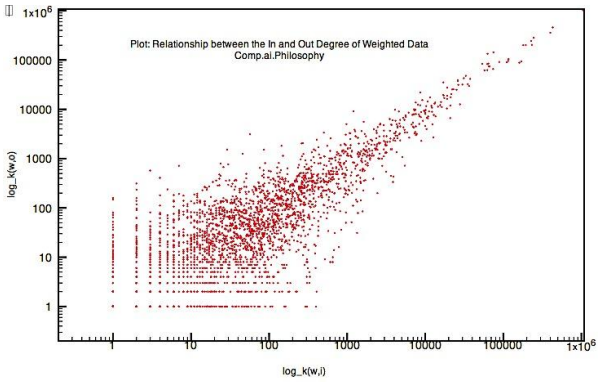


Figure 24 : Comp.ai.philosophy graphs for degree

5.2 Average shortest path length and diameter

The average shortest path length in both groups is relatively similar, when rounded up both have a shortest path length of three, which corresponds to any two randomly selected actors within the network are on average three shortest path lengths apart or can be reached through any two other actors. The fact that the two groups display similar shortest path lengths is unsurprising as both have a high number of connections for the average degree. This is also similar for that of the sample groups. However when compared to the average shortest path lengths of a random graph, the real data is much smaller, again this is due to the large number of edges for both. The diameter, which is the maximum shortest path length between any two actors, is also the same in both groups. Results are different to both the sample data and the random graph, which, in both it was expected the Tex group would have a large diameter than that of the AI group.

5.3 Centrality and Clustering coefficient

The philosophical group has a larger graph density and global clustering coefficient when compared to the Tex group. For a completely connected graph these two values should be equal to 1. Therefore the clustering coefficient alone shows a highly connected graph, and although the graph density is minute it is still larger than the Tex group showing the philosophical group to be more densely connected. Figure 26, show the number of actors with corresponding coefficients. The Tex group shows a high number of actors with a wide spread of coefficients

between the value of 0 to 1. The AI group shows a large selection of actors with values greater than 0.5 which would correspond to the high average clustering coefficient.

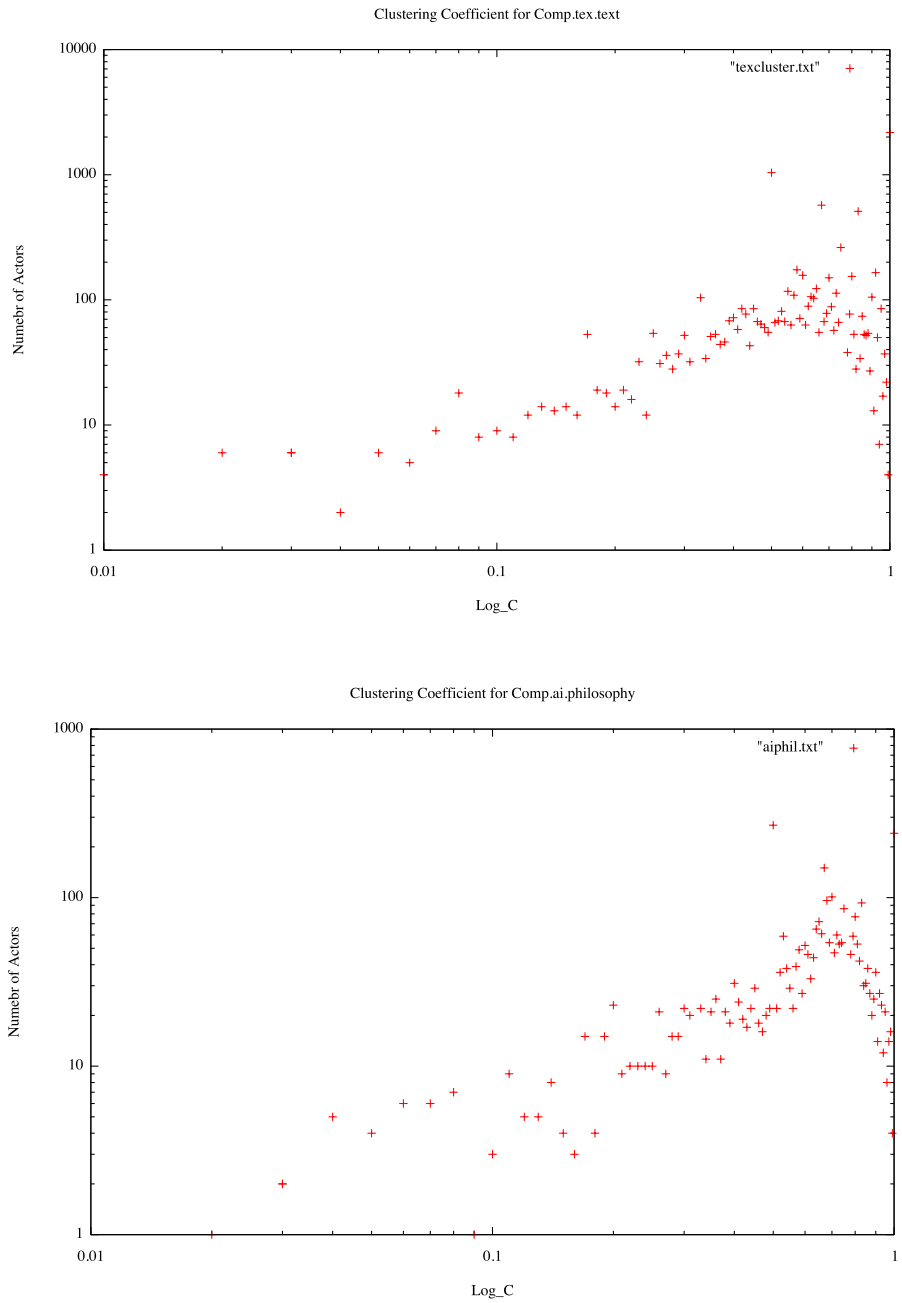


Figure 25 : Clustering coefficient

The random graph produces a clustering coefficient of 0.00516442 for the AI group and 0.000735558 for the Tex group. This is calculated using the formula $\frac{\bar{k}}{N}$ for the clustering coefficient of the random graph. This means both groups display high clustering when compared to the random graph of equal size N and same degree. These values shows that the nearest neighbours of any actor are likely to have replied to the same post if they have a 'friend' who has also replied to that post. This applies particular to the AI group. Due to this fact, discussion groups where a majority of actors display the question answer and spam roles are likely to have a lower global clustering coefficient than groups where a majority display the discussion role. The AI group has 89% of individual actors with 0.5 clustering coefficient or higher compared to only 52% of actors in the Tex group.

Although both density values are small, the AI group is higher than the Tex group, this may be due to the number of connections between two actors.

5.4 Betweenness

A measure from 0 to 1 where 1 informs that every actor within the network contributes equivalently to the connectivity between the other actors. An example is a graph formed by vertices connected in a single cycle A graph with a small average diameter and shortest path length is expected to show a high average

betweeness value. Following the results above, the betweenness value is expected to be similar in both groups.

The global betweenness for the comp.ai AI group is 0.000344183, and for the comp.tex.text group is 0.000070894, both have extremely low betweenness which is related to the relatively high clustering clustering coefficient. The last indicates that many vertices are directly connected without intermediate vertices. Therefore, only a minority of vertices provide connectivity that would contribute to betweeness.

5.6 Time Line

The time evolution of all actors in both groups is displayed in figure 27. The vertex degree of each actor grows over time. As there are many actors within both networks it is difficult to determine the time of individuals in both groups. However it is very clear that the degree of a selected few actors increases dramatically by a large amount over time.

This includes all the top posters seen in this report. High actors in the Tex group are continuing to post past the 7000 days. High actors in the AI group may leave the network at any time or are not continual actors up to 7000 day period of observation.

Photographs and full names have been removed from these tables for data protection reasons.

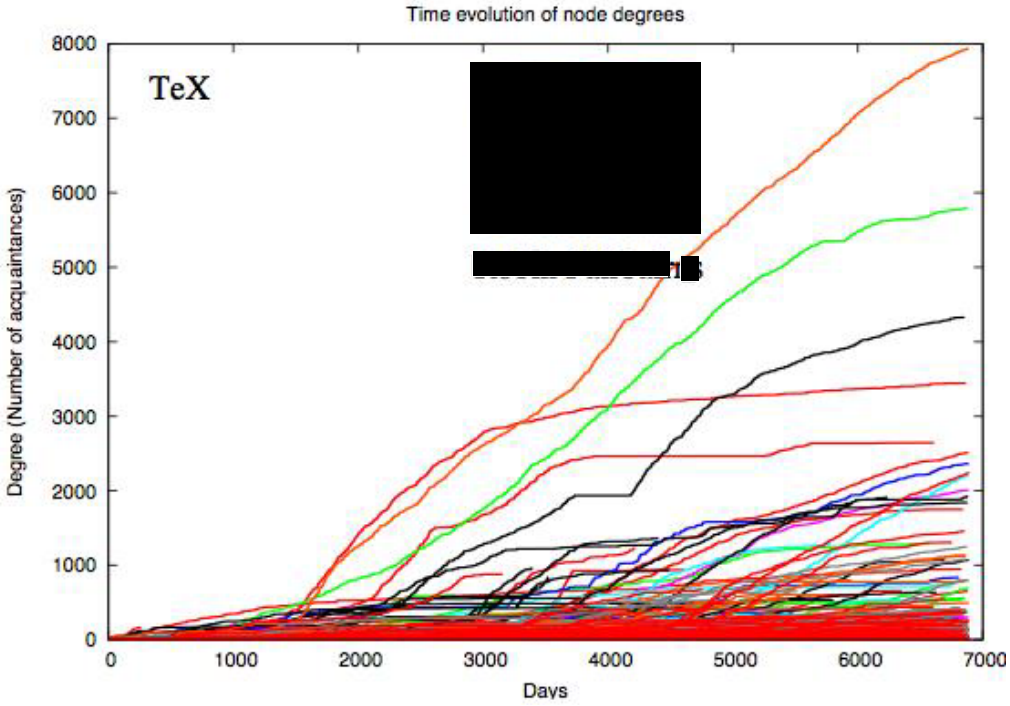


Figure 26 : Time evolution of actors in Tex group

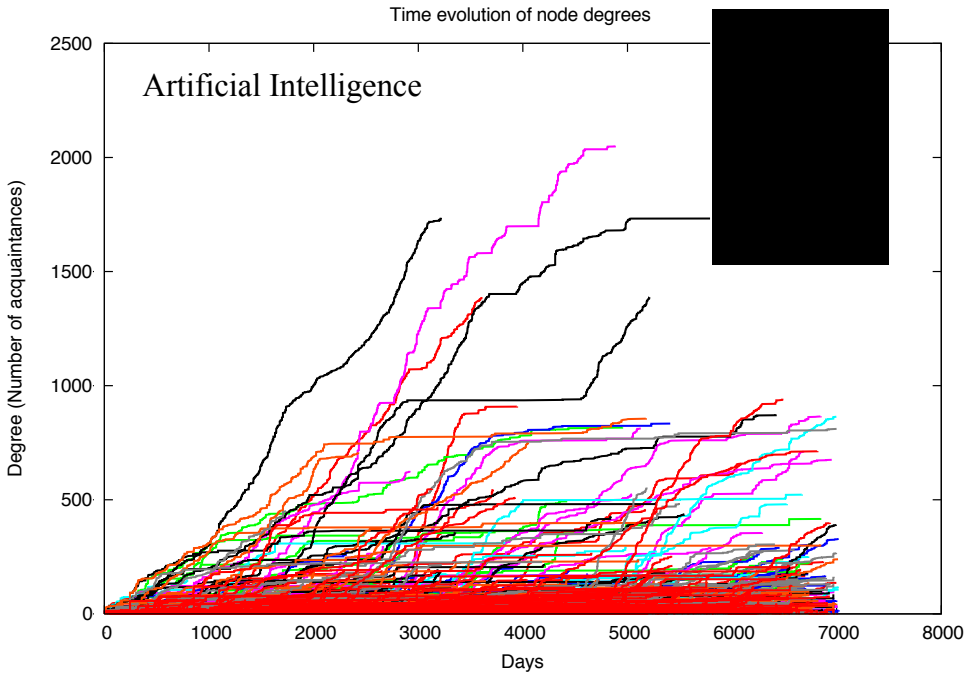


Figure 27 : Time evolution of actors in AI group

The average time of an actor in the AI at 57.5 days is greater than the Tex group at 2.43 days. The top actors of the both groups are not the longest members of the group.

5.7 Scale-Free Behaviour

This network does not follow the behaviour of a random graph, instead it shows a power law degree distribution and the evolution of the graph suggests that it is scale free.

In the nature of the AI discussion group where thread lengths, and number of posts are longer than the Tex group, an actor chooses what discussion topic to reply to or what post to start up. A post will display a number underneath the post which shows how many replies it has generated. An actor may read up on a post with a high number of replies to it as they may be interested in what causes this. This will be classed as preferential attachment and one may assume posts that have a high number of replies generate more replies through this attachment scenario.

For the Tex group this may not necessarily be the case as once a question has been answered there would be no need for further posts on the topic.

However one can not be certain how users with high degree have got this. They could simply have replied to only a few posts but these few posts may have a high number of replies creating a high degree. Or they may have replied to a lot of posts with few replies. Or they may simply have created a lot of posts and never replied.

A questionnaire to the users of this network might help understand what motivates a user to reply to a post.

Age may also limit the behaviour of scale-free as a user who logs onto the group today may not look back 10 years to find a post that has a high degree.

5.8 Summary

Although the diameter, and the average shortest path length did not provide any differences they were still relatively small when compared to the number of vertices and edges between. Other properties confirmed the hypothesis, comp.ai.philosophy has a higher degree and weighted degree, higher average clustering coefficient and a smaller betweenness when compare to comp.tex.text.

6.0 COMPARING POSTS POSITIONS

The sample data shows that the AI group should have a large number of posts in the middle positions when compared to the Tex group. For the Tex group one also expects to have majority of posts in the middle position, however due to shorter threads one may expect to find a wider spread with posts in the first and last position. Idle posts are expected to occur with frequency in the Tex group than that in the AI group. The latter is due to the amount of announcements of new releases of packages and updates of the Tex systems.

6.1 First Position

The number of posts in the first position of the thread is counted when the thread length (number of posts in a thread) is greater than one. Thread lengths of one are idle posts.

In figures 29, 30 and 31 we investigate the correlation between the overall number of posts of each actor in the two groups with their number of posts in an first, intermediate (middle) or final (last) position.

In these figures each actor is represented by a symbol. In figure 29 the coordinates of the position of the symbol are the overall number of posts (x-axis) vs. the number of posts in first position.

In figures 30 and 31 the y-axis coordinate counts the number of posts in the middle and last position.

Only one symbol is shown if several actors happen to have the same coordinates.

	Comp.ai.philosophy	Comp.tex.text
Average % of actors in first position	12.83	25.29

Table 13 : First position

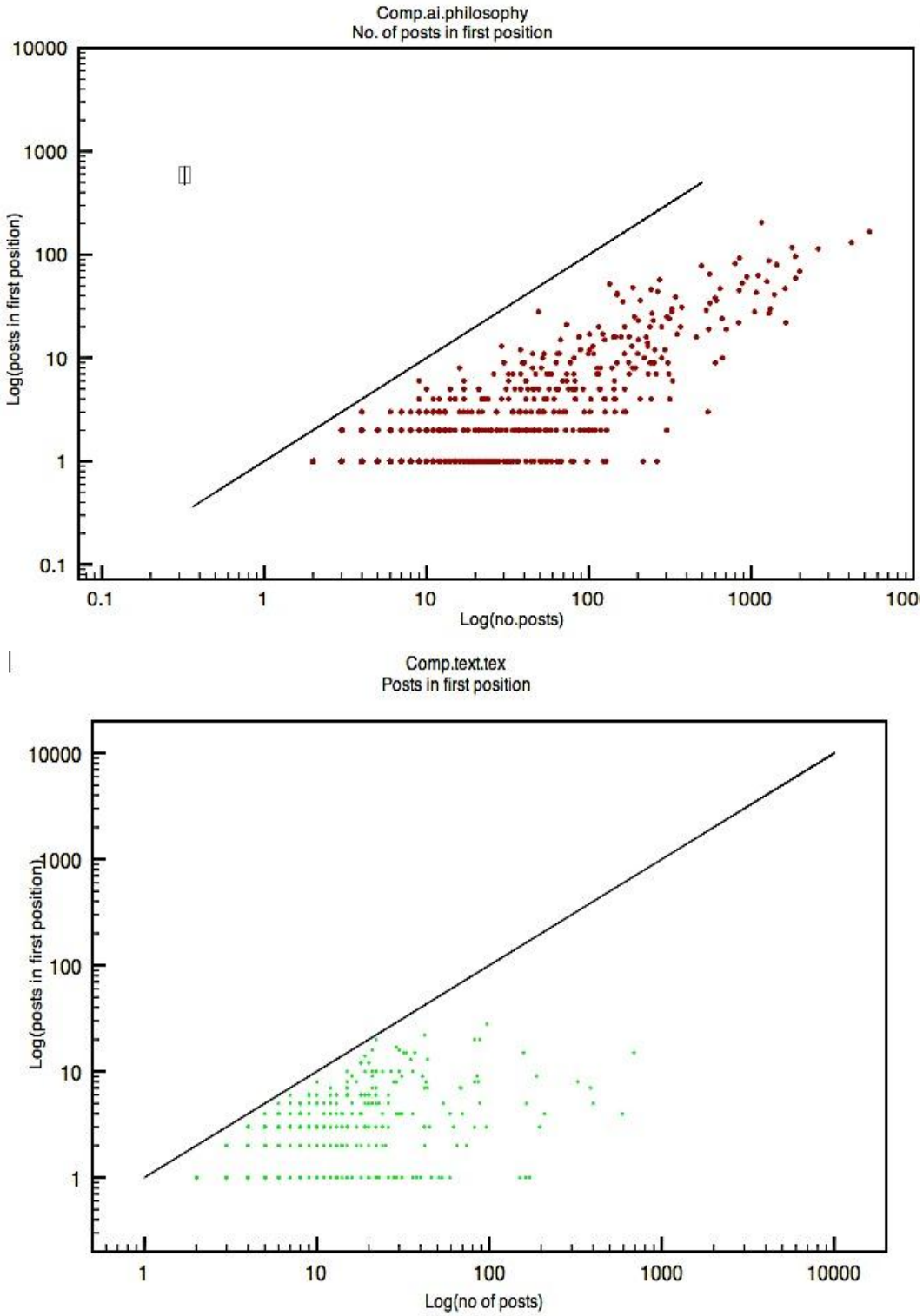


Figure 28 : Posts in first position

Figure 29 shows the number of posts in the first position relative to the number of posts for the actor. The line $f(x) = x$, shows a great percentage of their posts are in the first position, the actors who start a thread. On average 25% of a

Comp.text.tex groups posts are in the first position which exceeds 12% on average of Comp.ai.philosophy.

6.2 Middle Position

The middle position is calculated on threads with thread length greater than two, whose posts are not commencing and not concluding the thread.

	Comp.ai.philosophy	comp.tex.text
Average % of posts in middle position	42.97	25.32

Table 14 : Middle position

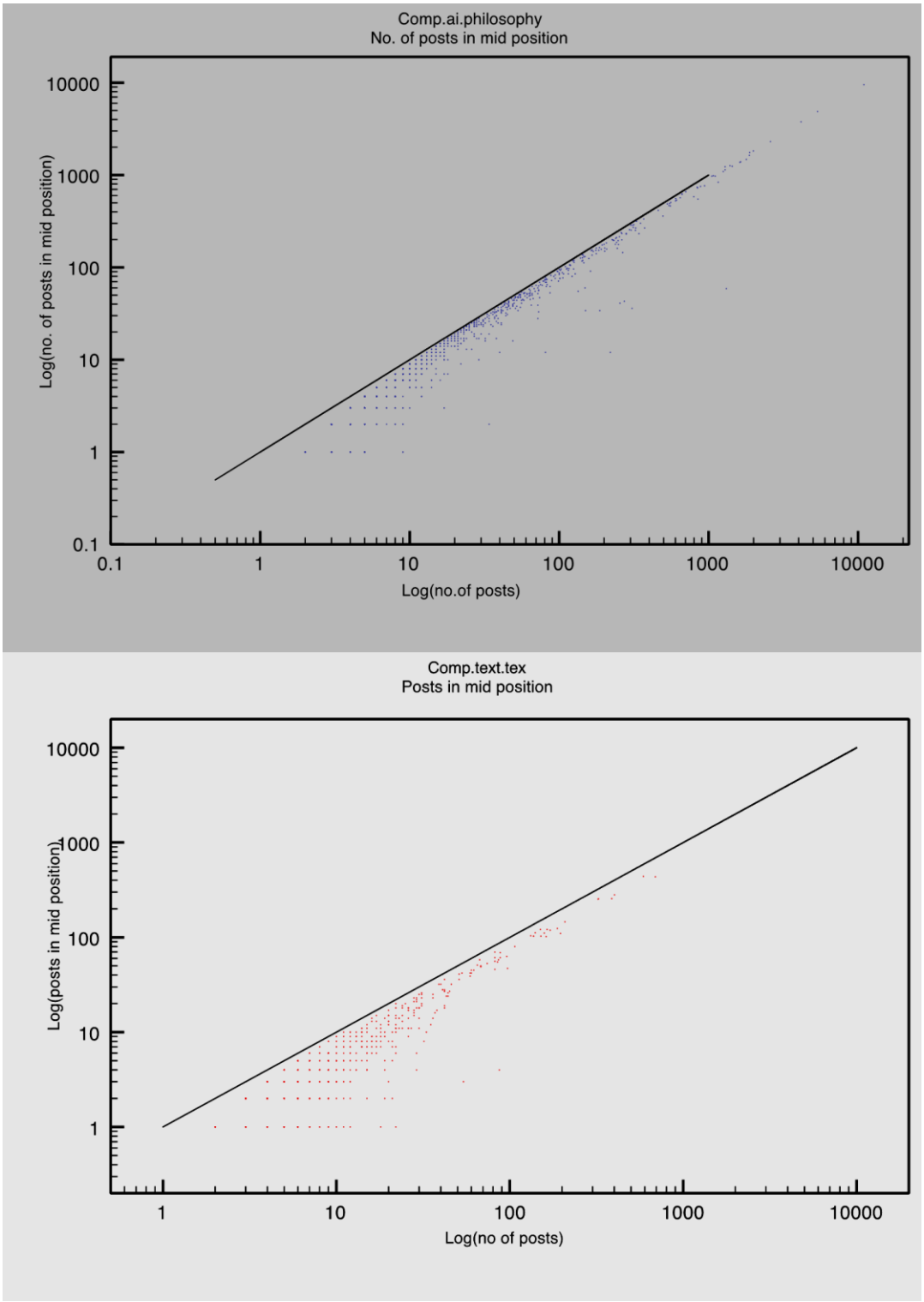


Figure 29 : Posts in middle position

The results in figure 30 confirm expectations that comp.ai.philosophy has a greater percentage of its posts in the middle position. In fact 43% of its posts are in the middle position far greater than 25% of posts for comp.text.tex group. This is due

to the nature of the group, as greater depth (higher average thread length and high reply degree) into discussion is present. From figure 30 one may also observes that the maximal number of middle position posts in the AI group exceeds by factor of 10 those of the Tex group. This corresponds to the much longer average thread length.

6.3 Last position

The last position is calculated from posts with thread length greater than one.

	Comp.ai.philosophy	Comp.tex.text
Average % of posts in last position	3.43	11.433

Table 15 : Posts in last position

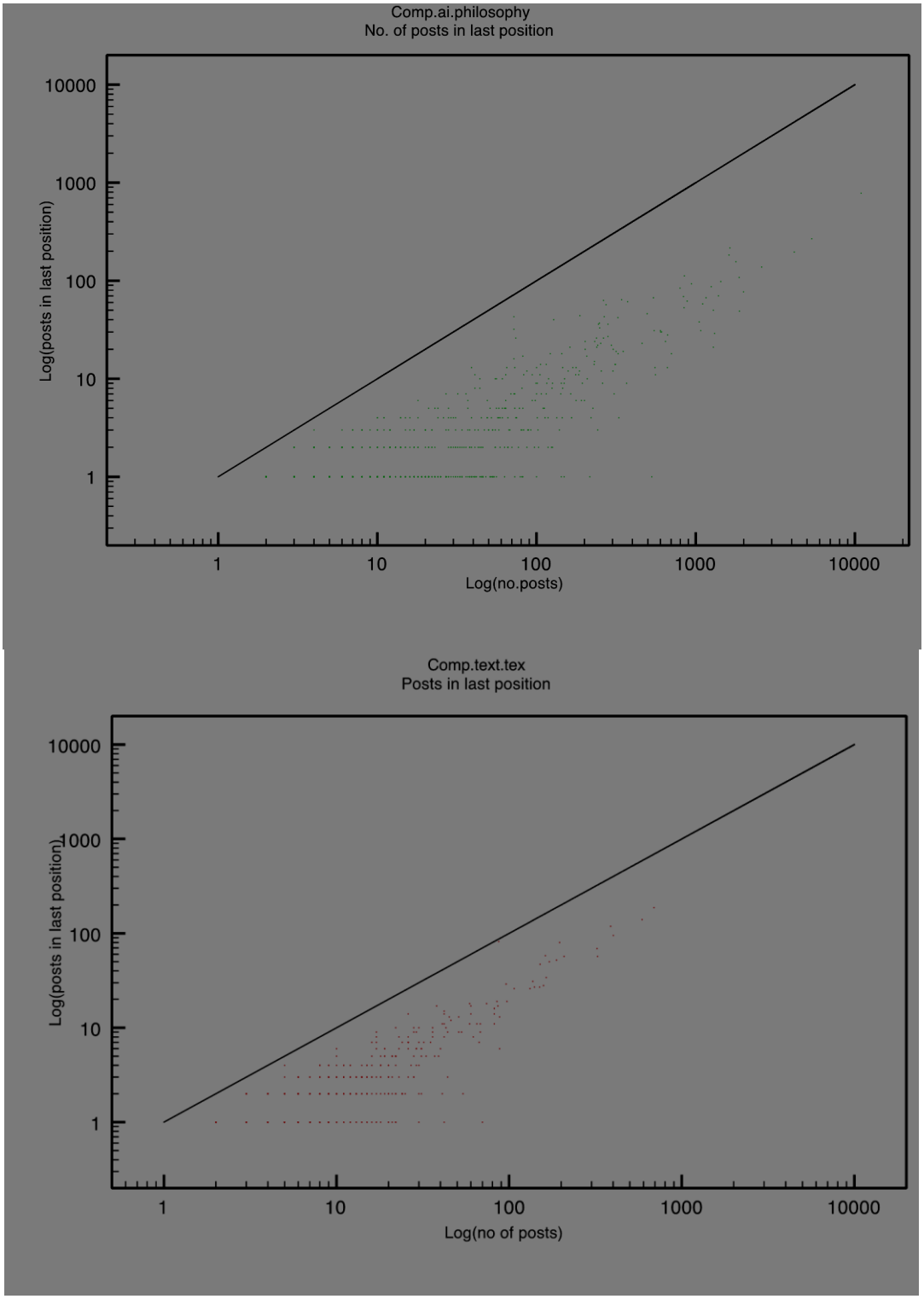


Figure 30 : Posts in last position

It is expected that actors with a high amount of their posts in the last position are answer people. On average comp.ai.philosophy has a small amount of posts in the

last position (3%). This is slightly mirrored in the results for comp.text.tex event, although it is greater than the comp.ai.philosophy group it still has on average a small percentage (11%) of a actors posts in the last position.

Overall comp.ai.philosophy has confirmed our hypotheses that a large amount of individual's posts are in the middle positions. Opposed to this comp.text.tex was expected to have less posts in the middle position compared to the first and last positions. This is not entirely what has happened here, actors seem to have a higher amount of posts in the first and middle positions compared to that in the final.

6.4 Comparing Threads Counts

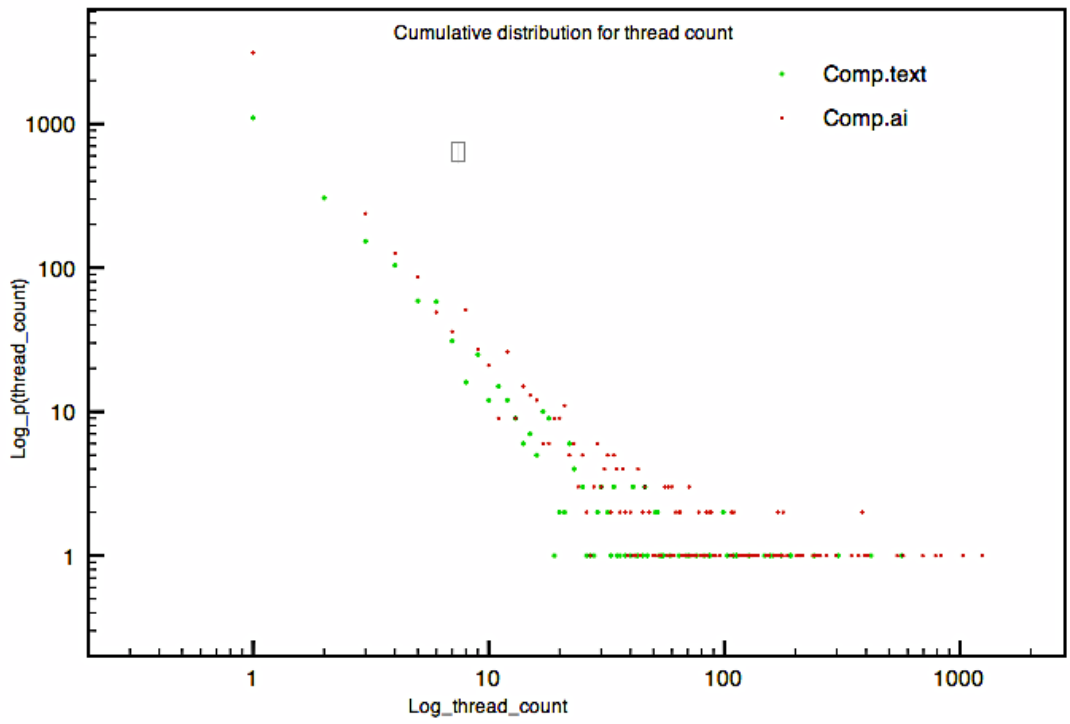


Figure 31 : Distribution of thread count for both groups

The above graph shows the distribution for thread count for both groups. The thread count is the number of threads an individual has posted to. Both groups show a similar distribution and as expected Comp.ai.philosophy group thread count is higher than Comp.text.tex group.

	Comp.ai.philosophy	Comp.tex.text
Max Thread Count	1920	565
Average Thread Count	6.858	5.276
Most common thread count	1	1

Table 16 : Thread count

Differences between the two groups can be seen above in the table, with maximum thread count for Comp.ai.philosophy group on average three times greater than Comp.text.tex group. Globally it is expected that actors of question and answer will only interact a small number of times, and hence should have a small thread count and vice versa for the discussion role. However the thread count does not display such vital information on it's own account. It is only when it is coupled with the thread length that such information can be obtained.

6.5 Thread Length

The thread length is the number of posts in a thread. Results for both groups can be seen below

- Comp.text.tex maximal thread length: 243
- Comp.ai.philosophy maximal thread length: 1347

Displayed in figure 33 is the cumulative thread length this shows that the AI group has up to ten times larger thread lengths than that of the Tex group. Although both thread lengths display a slow decay in the distribution, the AI group has larger thread lengths and a slower decline with $\gamma = 2.90$ compared to Tex group whose $\gamma = 3.35$. With these results and the thread count, the AI group actors are discussing more in one thread than that in the Tex group. This may cause further new threads to prosper. The thread count for the Tex group may differ from that of the AI group due to a lower number of posts per thread. Once a question is answered there is no need for further discussion. If there is no thread header asks or answers the question an actor will be inclined to start a new thread.

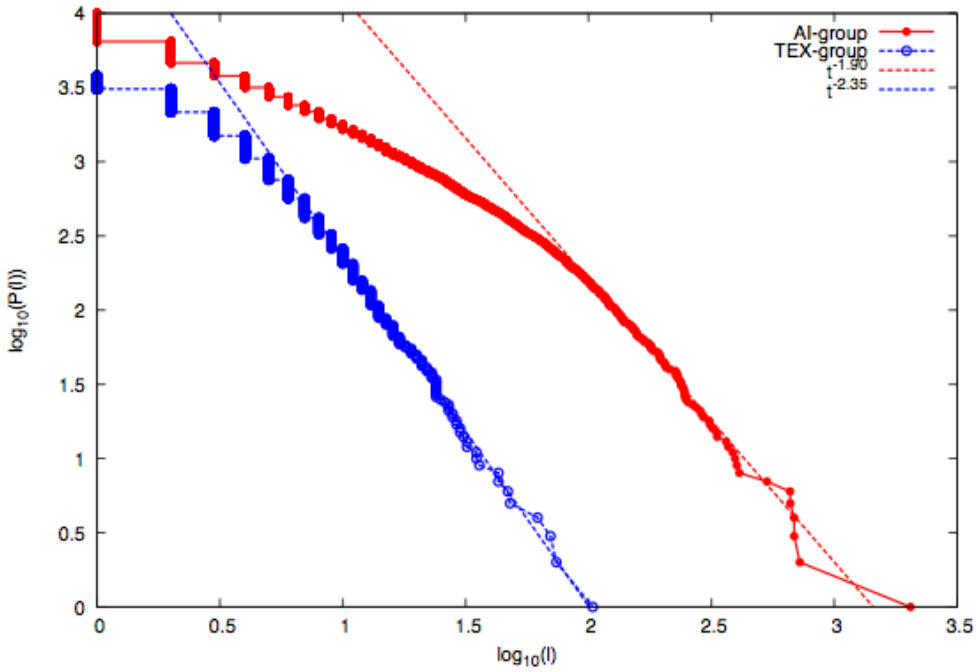


Figure 32 : Cumulative thread length for both groups

6.5.1 Idle Posts

A post is called idle when an actor of the network initiates a post and receives no reply. These could be posts in which a question has been asked but received no answer, a discussion started but received no reply or spam posts. Each of these posts have thread length one.

	Comp.ai.philosophy	Comp.tex.text
No. of Posts	73907	149654
No. of idle posts	5518	695
% of posts that are idle	7.47 %	0.46 %

Table 17 : Idle posts

Table 17 shows the number of idle posts in each group and confirms what can be seen in the cumulative thread length distribution. The AI group has a greater share idle posts with 7.47% of the overall posts remaining idle. The Tex group has a minute amount of its posts being idle with only 0.46%.

There are a total of 1344 actors counting only actors with idle posts giving an average number of idle posts per actor of 4. However the Tex group has a greater average of 15 idle posts per such actor because the number of actors with idle posts is considerably smaller with only 45 actors.

Comp.ai.philosophy	% of idle posts	% of total posts
BG01	21.64	1.61
AL01	4.6	0.34
AW01	3.5	0.26

Table 18 : AI group idle posts

Comp.tex.text	% of idle posts	% of total posts
AR01	9.21	0.43
FR01	7.48	0.035
HJ01	6.19	0.028

Table 19 : Tex group idle posts

Table 19 shows the top three actors of idle posts, unsurprisingly the maximum of the idle posts for the AI group is greater than that of the Tex group. By considering the total of the top three idle actors the AI group is ten times greater than the Tex group.

The AI group's top three idle actors contribute to a large per cent, 29.1% of the total number of the idle posts and to 2.21 % of the total number of the posts.

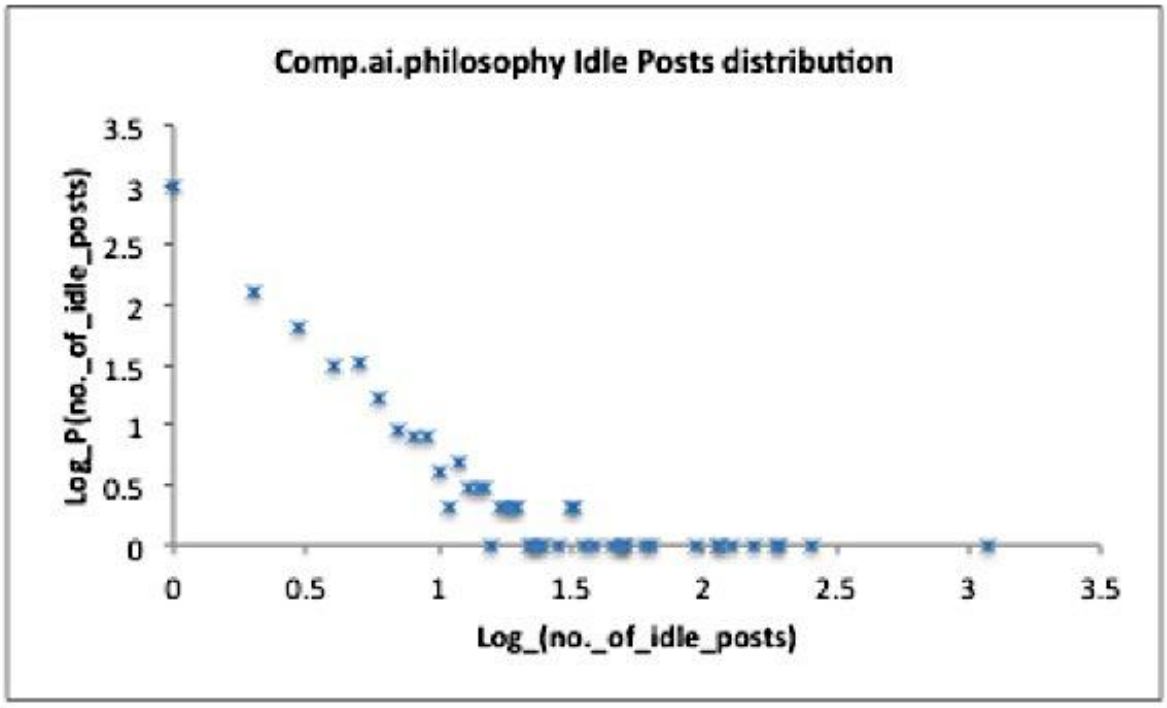


Figure 33 : Idle posts distribution for AI group

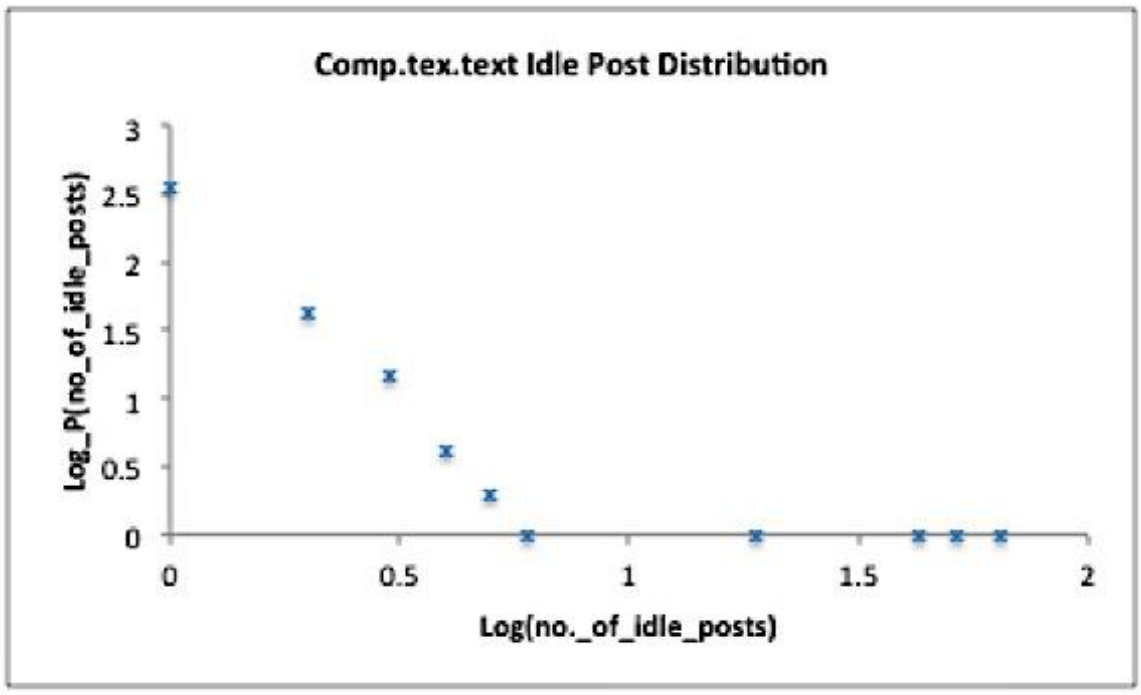


Figure 34 : Idle post position for Tex group

The graphs displayed in figures 34 and 35 show the distribution of the number of actors with idle posts. The Tex group shows a sharper decline in the data and a

small range of values. One of the explanations for these differences is that in the Tex group there are a small number of posters who announce new Tex features or packages – posts that do not ask for any answers.

Within these idle posts there are actors of each group, which post only idle posts and have never received a reply to any post they have initiated. The table 19 summarises these results.

6.5.2 Threads of length two and three

Threads of length two or three are expected to consist of questions and answer post for length two, or a question and answer and a further answer for post three. Therefore it is expected that the Tex group will have a higher amount of threads with length two than that of the AI group. The table below shows the results for each group.

	Comp.ai.philosophy	Comp.tex.text
Total Number of Actors	3783	14264
Actors with Idle posts	1344	425
Idle Actors	924	148

Table 20 : Idle actors only

As expected the AI group has a greater number of actors whom only have idle posts compared to the Tex group. 68% of the actors with idle posts have only idle posts, with 34% of the Tex group having idle posts. Of the total number of actors

24.4 % of the AI actors have only idle posts which is higher when compared to the 1.04% of the Tex group.

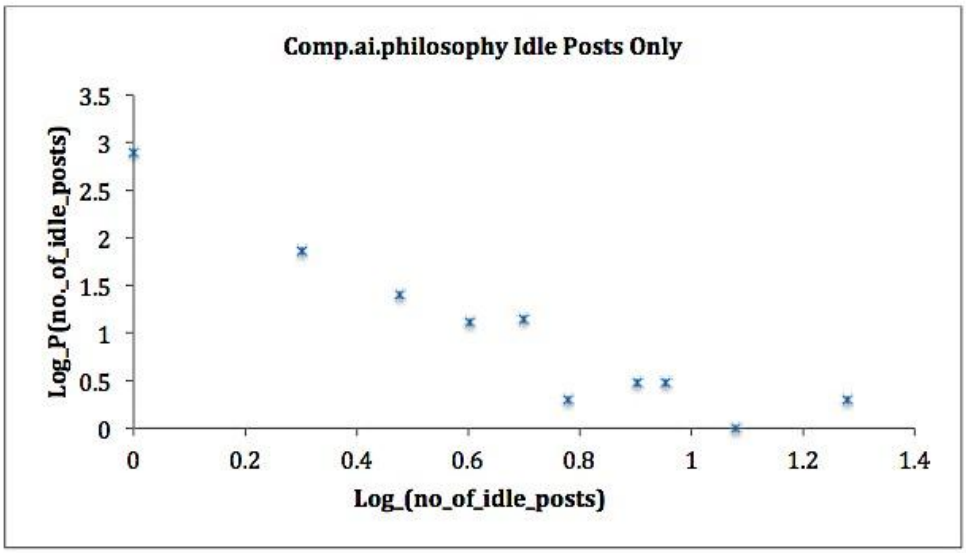


Figure 35 : AI idle posts

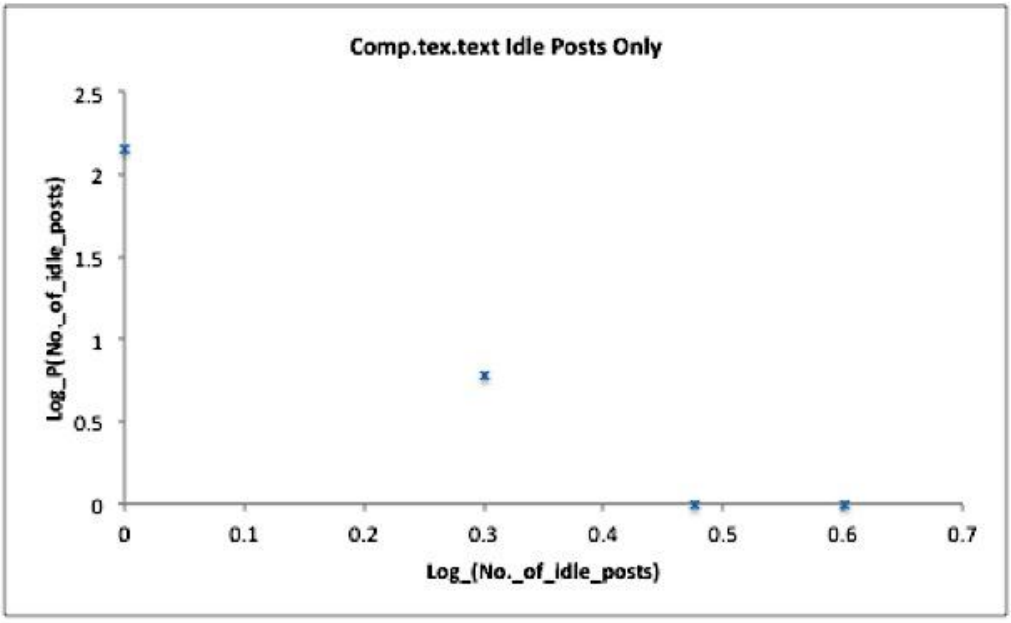


Figure 36 : Tex idle posts

The graphs display similar properties; there is a high number of actors with only one idle post and small number of actors with many idle posts. The difference between the two distributions is the highest number of idle posts in the Tex group

is much smaller than the highest in the AI group. Also in the AI group there is more of a spread in data of the actors with only idle posts.

	Comp.ai.philosophy	Comp.tex.text
Total number of threads	11199	40722
Length 2	1690	82031
Length 3	794	6122

Table 21 : Threads of length two and three

For both thread lengths of two and three the Tex group has a combined percentage of 35% of its threads with these lengths, which is a large percentage when compared to the AI group with only 22%. This also shows that the AI group has a higher number of threads with length four or more, with this and the high number of posts this confirms more interaction between actors in fewer threads.

The table below provides the names of the most common actors with thread length two. To find if there is a relationship between any two actors, if actor a posts and actor b replies, then if actor b starts a post and actor a replies. Combinations of the two threads have been combined to give the below results.

Comp.ai.philosophy	
ZL01 – ZL01	58
BG02 – BG02	39
BG03 – BG03	37
Comp.tex.text	
TC01 – CT01	173
LG01 – OH01	23
LG01 – LG01	*13*

Table 22 : Relationships between actors of thread length two

The Tex group provides interesting results with a large amount of the threads contributed by the relationship between TC01 and CT01. TC01 is in fact an automated system message announcing updates on software. There are then a high number of posts between LG01 and OH01, and the third most common thread length two is a self reply relationship between LG01 and themselves. There are in total 736 threads of length two that are self replies in the Tex group, which contribute to a small 8% of the entire threads. This means that there is a high per cent of threads where different actors provide the first post and the second.

The AI group's most common actors of thread length two are all self replying threads, where for the whole group 442 of its threads of length two are self replies, which provides 26% of the total. Table 23 then provides the three relationships of users that are not self-replies. These do not provide a great amount of contribution to the number of thread lengths of two.

IG01 – PJ01	9
CA02 – ZL01	9
SH01 – AR02	8

Table 23 : AI group top relationships of thread length two that are not self replies

The top relationship in threads of length three are provided in table 24. Again, if an actor *a* initiates a post, and actors *b* and *c* then reply, or if actor *b* starts a post and actors *a* and *c* reply, etc. These thread lengths are combined to find a relationship between three actors.

Comp.ai.philosophy	
BG02 – BG02 – BG02	7
IG01 – BJ01 – TT02	6
YC01 – YC01 – YC01	6
Comp.tex.text	
RN01 – MN01 – HE01	47
LG01 – OH01 – LG01	12
Randy Yates – VH01 – Randy Yates	5

Table 24 : Top three way relationships for thread length three

The top three relationships of the AI group provides a small number of the total threads, informing that there are not too many three way relationships formed in the AI group.

The top relationship in the Tex group are spammers, this is known as these were present in the sample group of the network and hence a sample of their posts were read. Their posts consisted of advertising websites. There is a small mixture of three way relationships within both groups, consisting of self-replies, which are threads of length three but only one actor. There are also threads of length three that have only two actors, and interestingly the Tex groups second top relationship is between LG01 and OH01 who were present in the thread length two results. There are also relationships between three different actors.

6.6 Summary

In summary, the thread lengths investigation show that the AI group has a greater number of idle posts and actors who only have idle posts. It also has a small number of threads of length two or three while on average the threads are ten times longer that the Tex group. With all these results, the AI group is predominantly clasiified as a discussion group based alone on the investigation of thread lengths.

The Tex group has a small number of idle posts and a great number of threads with length two and three. There is also a small amount of self replies proving more relationships to be formed between any two actors. A technology group where question and answer posts are expected to dominate should not need lengthy threads and this group confirms this relationship.

7.0 INDIVIDUAL ACTORS

As the global statistical properties have been investigated it would be wise, as in the sample group, to look at top actors in each of these groups and see if they display any properties of any of the roles.

All data and names within this group are available to any member of the public, however names have been replaced by alternative codes to protect the identity of the individuals. It was assumed that two actors with the same email address are the same actor and hence these posts were combined.

7.1 Comp.ai.philosophy

The results of this group have shown a high number of actors displaying the discussion role . In the following section the top actors in table 25 and table 26 are examined individually as separate networks in which all actors are connected to the top actor.

The in and out-degree for each actor is relatively proportional to each other as it is expected to be. The clustering coefficient for such a role is expected to be high as it was found to be in the sample and global properties of this network. The results of the individuals do not show this, this may be because each actor is of high degree and it is increasingly difficult to have every actors neighbour connected to one another.

The average shortest path length of a discussion role is expected to be small and the results confirm this. All of the top actors average shortest path length is very similar to the average shortest path length of the global network.

The average post length and thread count is also high for each actor.

Actor	ZL01	WC01	KW01	SG01	LD01
Degree	2238	1850	1460	1379	1110
In-degree (%)	46.16	46.38	48.08	44.89	46.76
Out-degree (%)	53.84	53.62	51.92	55.11	53.24
Weighted Degree	2077154	379969	436198	221875	395459
Weighted-in-degree (%)	50.54	47.54	53.31	527.22	49.11
Weighted Out-degree (%)	49.46	52.61	46.69	47.28	50.89
Clustering Coefficient	0.022	0.025	0.035	0.036	0.041
Average Shortest Path Length	2.101	2.165	2.101	2.107	2.116
Reply Count	12.81	10.5	3.21	4.69	9.90
Thread Count	10818	5314	1970	2563	4107
Average Post Length	55.6	125.781	41.65	95.89	80.49

Table 25 : AI Top 5 Actors

Actor	FR01	KD01	ML01	FU01	AD01
Degree	6362	3072	4101	4132	3627
In-degree (%)	40.21	44.11	25.04	43.37	64.19
Out-degree (%)	59.79	55.89	55.45	56.63	35.81
Weighted Degree	85506	82662	39863	39706	34895
Weighted-in-degree (%)	46.51	48.20	53.99	50.46	42.27
Weighted Out-degree (%)	53.53	51.80	46.01	49.54	57.73
Clustering Coefficient	0.001	0.005	0	0.002	0.003
Average Shortest Path Length	2.221	2.115	2.157	2.143	2.176
Reply Count	3.92	7.7	8.57	5.96	4.61
Thread Count	637	320	185	207	584
Average Post Length	28.37	31.53	41.81	30.652	19.93

Table 26: Tex Top 5 Actor

ZL01 is prominent in both the weighted and non-weighted degree, and is the highest posting actor in this group with a ratio of weighted in to out-degree at 1.02 and non-weighted 0.86. ZL01 hold's a small clustering coefficient, and average shortest path length. The thread count is the largest of the top actors with a relatively small average post length. The reply count is large with 12.8% of its connections as self loops.

WC01 is the actor with second highest post count in this network with 5348 posts, and has a similar in and out-degree. Results such as clustering coefficient and average shortest path length are similar to ZL01. WC01 has longer posts and is involved in less threads than ZL01

The third top actor for the comp.ai.philosophy group yet the lowest degree shows that although LD01 has been involved in 4107 threads and the average thread length is fairly long at 80.46, they do not seem to be connected to the same amount of actors that other are in the total group. This may be because they post to threads that only contain a small amount of users. This causes the clustering coefficient to be the greatest within the top actors. LD01 has a similar average shortest path length to other actors.

The next top actor SG01 similar to all other actors, has a small clustering coefficient, and average shortest path length. LD01 posts overall 2592 times and with a thread count 2563 this corresponds to just over 1 post per thread.

The fifth top actor of this network is KW01, who has a similar in and out-degree small and clustering coefficient. The reply count is the smallest of the top actors.

They are involved in 1970 threads with 1994 posts, the posts per thread is 1.012 and average post length is the smallest out of the top actors at 41.65

7.2 Comp.tex.text Top Actors

Unlike the AI group where there were different actors for the top degree for both weight and non-weighted data, the Tex group has the same actors in each just in different order. The clustering coefficient varies for each of the actors from a zero to 0.005 which is extremely small.

FR01 is the highest posting actor for the Tex group for both the weighted and non-weighted data. The out-degree is greater than the in-degree. A result of the degree causes a small clustering coefficient, betweeness and average shortest path length. FR01 is not only involved in a large amount of threads but also has a large average thread length.

The second highest weighted actor within the group and the lowest non-weighted degree within the top five actors, KD01 displays connections to a greater number of actors. The out-degree is always larger than the in-degree showing that KD01 has answered more posts than they have posted. Although from within the top five actors KD01 has the highest clustering coefficient, although still relatively small. KD01 also displays the smallest average path length and betweeness out of the actors. With an average thread length of 31.53, KD01 has been involved in 320 threads and has a large amount of posts in which they have replied to more than once.

The third top actor in the comp.tex.text network is ML01. They not only have the third largest degree, but also the third largest weighted degree. There is a great difference between the weighted degree of FR01 and KD01 to ML01. Interestingly ML01 has a greater weighted in-degree over weighted out-degree suggesting they have more people reply to their posts over ML01 replying to other posts. However for the non-weighted degree the in-degree is less than the out-degree. This suggests that although ML01 has a greater number of replies, these replies could be from the same actors.

ML01 has a zero clustering coefficient which would suggest that their local neighbourhood is sparse with fewer connections between other actors. Although ML01 is involved in the least number of threads, the thread average length of these is the largest.

The fourth greatest actor of the weighted degree and the second actor of the non-weighted degree this shows that FU01 has more connections to individual actors. Their weighted in-degree is similar to the out-degree, however for the non-weighted results the out-degree is greater than the in-degree, showing that the posts FU01 is replying to are to more individuals than they are receiving posts from. Clustering coefficient, average shortest path length, and betweeness give similar results to other members of the top actors. The thread count is large with a corresponding large average shortest path length.

AD01 in the final position of the top actors with the least weighted-degree and the fourth non-weighted degree. The weighted out-degree is greater than the in-degree, however the non-weighted results are opposite to this showing almost

double the amount of the out-degree for in-degree. There is a large difference in-degree between FR01 and AD01 however this does not reflect on the results for clustering coefficient, betweeness or average shortest path length which are all small. The large thread count will contribute to the smallest thread length.

7.3 Post's positions of the top actors.

	First	Middle	Last	Idle
ZL01	4.50	87.31	7.14	1.03
WC01	3.12	91.23	5.01	0.62
KW01	3.46	91.47	3.86	1.15
SG01	4.40	89.16	5.32	1.08
LD01	3.15	90.93	4.71	1.18

Table 27 : Post positions of AI group

As a predominant discussion group whose both global results and sample results show that most posts are in the middle position of a thread, investigation into the posts positions of these top actors is necessary. Results of which are given below.

From these results, it is noticeable that each of these actors have an extremely large percentage of its post's in the middle position with every actor having greater than 80% of its posts in the middle position.

	First	Middle	Last	Idle
FR01	2.13	62.04	27.1	8.83
AD01	0.68	74.70	23.77	0.97
KD01	2.46	78.46	17.54	1.24
FU01	1.92	70.19	27.44	0.00
ML01	4.79	65.96	27.66	1.17

Table 28 : Post positions of Tex group

It can be seen that all of the top actors have majority of their posts in the middle position, followed by the last position and a small per cent in the first position and idle posts. Although this does not follow what was expected for this network, results for middle position are not as high as they are for AI group. The last position is also higher than the AI group.

7.4 Other Social Roles

It is clear from all the above results that AI group involves a large amount of actors who are discussion role. This does not mean that other actors will show roles of question, answer or spammer. Using results obtained from the sample group and primarily looking at the posts positions the table no of actors are examples of actors within this group displaying alternative roles.

7.4.1 Question Role

For both groups, to distinguish an actor that displays the role of the question actor they should display a high in-degree over out-degree for both weighted and non-weighted data. The non-weighted ratio to weighted ratio will be small and the posts positions will have a large percentage in the first position. All other properties are then calculated.

Actor	AH01	LS01
Degree	58	68
In-degree (%)	93.10	98.53
Out-degree (%)	6.90	1.47
Weighted Degree	133	360
Weighted-in-degree (%)	93.98	99.72
Weighted Out-degree (%)	6.02	0.28
% posts in First Position	90.9	55
Clustering	0.411	0.477
Thread Count	21	5
Post length	23.27	119.66

Table 29: Actors that display question role

The Comp.text.tex group AH01 has 90% of its posts in the first position, while SL01 from the Comp.ai.Philosophy group only has 55%. This is caused by the nature of the group. Both have a high in-degree, zero betweeness and similar clustering coefficient. AH01 is engaged in 21 threads of which 20 of these AH01 has initiated, and SL01 has only participated in 5 threads . They both have a reply count at zero and the major difference is the average length of one post, with AH01 being a lot smaller than SL01. SL01's average post length exceeds that of the average for the entire group. Most of these properties are typical of the role of a question actor.

7.4.2 Answer Role

The actors displaying the answer role have a greater out-degree over in-degree, the ratio of non-weighted to weighted degree is low and majority of its posts are in the last position.

Actor	HW01	HM01
Degree	666	75
In-degree (%)	36.64	4
Out-degree (%)	63.36	96
Weighted Degree	8567	574
Weighted-in-degree (%)	28.84	0.52
Weighted Out-degree (%)	68.01	99.48
% posts in Last Position	68.75	66.67
Clustering	0.116	0.365
Thread Count	10	5
Post length	34.8	36.1

Table 30 : Results of actors who display answer role

HW01 is an actor from the Comp.tex.text group and HM01 is an actor from the Comp.ai.philosophy group. As expected from results of the sample group, the clustering coefficient and betweeness values are small, although HW01 is smaller than HM01. Although the thread count differs greatly the average post length is similar in both groups. HW01 is only involved in ten threads and considering they have only ever posted sixteen times with a reply thread count at 2.3%, suggests that this reply could correspond to a number of posts in one thread. HM01 has only posted 6 times and been involved in 5 threads which is in line with the results.

7.4.3 Discussion Role

An actor who displays the properties of a discussion role must have a similar in and out-degree for both weighted and non-weighted. The ratio of weighted to non-weighted is high and there are not only a large amount of posts but also these posts are in the middle position.

Actor	SW01	TT01
Degree	876	423
In-degree (%)	45.09	51.77
Out-degree (%)	54.91	50.59
Weighted Degree	5538	139450
Weighted-in-degree (%)	60.80	53.24
Weighted Out-degree (%)	38.99	47.11
% posts in Middle Position	81	91.54
Clustering	0.0068	0.112
Thread Count	149	1059
Post length	22.27	40.52

Table 31: Actors who display discussion role

Although not all results above suggest that SW01 from the Tex group is a discussion role, for example low clustering, SW01 is the best candidate from the Tex group, of users that had not already been previously explored. TT01 from the AI group has greater values for the properties in table 31

7.4.4 Summary

Properties of individuals for each role type form similar results. A few differences are observed such as the average shortest path length for the answer role of the AI group is greater than the Tex group. This is due to the nature of the AI group which has been seen to show greater average shortest path lengths.

CONCLUSION

The main objective of this thesis is to classify roles within two very different discussion groups available online through Google groups.

It was expected that the AI group would shows actors to be predominantly discussion role's. With a high average degree for weighted and non-weighted data, a high clustering coefficient and small diameter, average shortest path length and betweeness. When these results are compared to the Tex group it confirms the expected results. A vast amount of posts are in the middle position, with similar in and out-degree and the average time an actor is a member of the group, all help to solidify this.

Individual actors in the AI group were investigated, with high actors showing the discussion role. The question role and answer role are also explored, although it was difficult in finding the actors who had a sufficient number of posts to provide data.

The Tex group was expected to have a large amount of question and answer users. Global properties of the group confirmed to have a higher number of vertices and edges than the AI group, low clustering coefficient, betweeness and average shortest path length. Positions of the posts were expected majority be in the first and last position. Results from both the sample and entire network show that a large number of posts were in the middle position. However the first and last

position percentage is greater than the AI group. Average thread length is ten times smaller than the AI group and the average time of an actor is considerably small at 2.43 days.

Overall the actors within the AI group show to be involved in long lengthy, in-depth threads. The Tex group show to be involved in short quick threads.

Although scale free behaviour is present due to the corresponding power-law distribution, how users are replying to posts is not yet determined. A further study would need to be done, and as previously mentioned a questionnaire asking how users choose to reply on a post could answer this.

This research provides areas for further study. The time line is only investigated as a global measurement. It would be interesting to assess the time line of top actors and individuals of the three roles. Maybe an actor would initially be classed as a question role and over time increase to an answer role. The role of an 'expert' would also prove to be interesting. What would class an actor as an 'expert', would time be a factor, number of posts, and degree.

Further investigation into the presence of capital words, (assumes an actor is shouting), question marks and famous philosophers, in an actors post could also provide further information in the role of an actor. As the Google groups website itself is a network a whole study on the whole network could also be investigated to see whether users are not only just sticking to one discussion group but are a part of the whole Google groups. To see what discussion topic is most popular and

also seeing whether there are certain individual users who participate majorly in many discussion topics are key to keeping the Google groups website growing.

As can be seen there are many investigations that can be created from this report, with more time available more results could be found. This type of investigation can be used on any other social or online social network.

REFERENCES

Adamic, Lada A., Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. "Knowledge sharing and yahoo answers: everyone knows something." In *Proceedings of the 17th international conference on World Wide Web*, pp. 665-674. ACM, 2008.

Albert, Réka, and Albert-László Barabási. "Statistical mechanics of complex networks." *Reviews of modern physics* 74, no. 1 (2002): 47.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. "Gephi: an open source software for exploring and manipulating networks." In *ICWSM*. 2009.

Boccaletti, Stefano, Vito Latora, Yamir Moreno, Martin Chavez, and D-U. Hwang. "Complex networks: Structure and dynamics." *Physics reports* 424, no. 4 (2006): 175-308.

Chang, Chin-Lung, Ding-Yi Chen, and Tyng-Ruey Chuang. "Browsing newsgroups with a social network analyzer." In *Information Visualisation, 2002. Proceedings. Sixth International Conference on*, pp. 750-755. IEEE, 2002.

Donath, Judith, Karrie Karahalios, and Fernanda Viegas. "Visualizing conversation." *Journal of Computer-Mediated Communication* 4, no. 4 (1999): 0-0.

Dorogovtsev, Sergey N., and Jose FF Mendes. "Evolution of networks." *Advances in physics* 51, no. 4 (2002): 1079-1187.

Fisher, Danyel, Marc Smith, and Howard T. Welser. "You are who you talk to: Detecting roles in usenet newsgroups." In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, vol. 3, pp. 59b-59b. IEEE, 2006.

Freeman, Linton C. "Centrality in social networks conceptual clarification." *Social networks* 1, no. 3 (1979): 215-239.

Gleave, Eric, Howard T. Welser, Thomas M. Lento, and Marc A. Smith. "A conceptual and operational definition of'social role'in online community." In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, pp. 1-11. IEEE, 2009.

Hanneman, Robert A., and Mark Riddle. "Introduction to social network methods." (2005).

Maia, Marcelo, Jussara Almeida, and Virgílio Almeida. "Identifying user behavior in online social networks." In *Proceedings of the 1st workshop on Social network systems*, pp. 1-6. ACM, 2008.

Mislove, Alan, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. "Measurement and analysis of online social networks." In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 29-42. ACM, 2007.

Molloy, Michael, and Bruce Reed. "A critical point for random graphs with a given degree sequence." *Random structures & algorithms* 6, no. 2-3 (1995): 161-180.

Newman, Mark EJ. "The structure and function of complex networks." *SIAM review* 45, no. 2 (2003): 167-256.

Newman, Mark EJ, Duncan J. Watts, and Steven H. Strogatz. "Random graph models of social networks." *Proceedings of the National Academy of Sciences of the United States of America* 99, no. Suppl 1 (2002): 2566-2572.

Panzarasa, Pietro, Tore Opsahl, and Kathleen M. Carley. "Patterns and dynamics of users' behavior and interaction: Network analysis of an online community." *Journal of the American Society for Information Science and Technology* 60, no. 5 (2009): 911-932.

Turner, Tammara Combs, Marc A. Smith, Danyel Fisher, and Howard T. Welser. "Picturing Usenet: Mapping Computer-Mediated Collective Action." *Journal of Computer-Mediated Communication* 10, no. 4 (2005): 00-00.

Welser, Howard T., Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. "Finding social roles in Wikipedia." In *Proceedings of the 2011 iConference*, pp. 122-129. ACM, 2011.

Welser, Howard T., Eric Gleave, Danyel Fisher, and Marc Smith. "Visualizing the signatures of social roles in online discussion groups." *Journal of social structure* 8, no. 2 (2007): 1-32.